

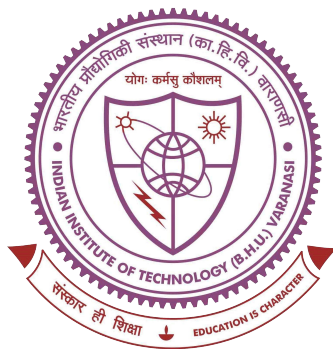
Sentiment Analysis – Disaster Tweets Classification (U.G. Project)

Mr. Piyush Pransukhka


Mathematics and Computing
Indian Institute of Technology (BHU)
Varanasi - 221 005

Mr. Shrey Gupta

Mathematics and Computing
Indian Institute of Technology (BHU)
Varanasi - 221 005



Outline

- Introduction and Motivation
 - Binary Classification
 - Sequential Models
 - Recurrent Neural Network (RNN)
 - Gated Recurrent Unit (GRU)
 - Long Short Term Memory (LSTM)
 - Tokenizer
 - Word Embeddings
 - Exploratory Data Analysis
 - Creating Training Data
 - Creating the Model
 - Results on Training Data
 - Results on Testing Data
 - Conclusion
 - Improvements and Future scope
- 

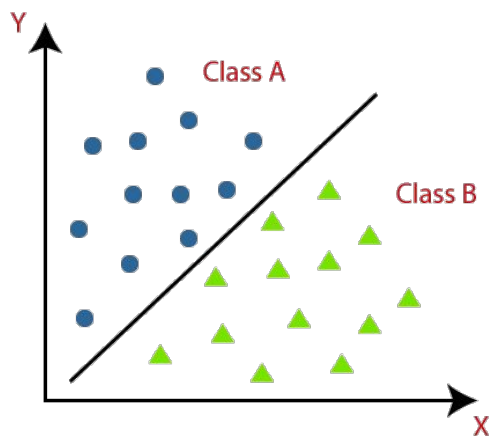
Introduction and Motivation

- Sentiment analysis is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text.
- Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. But, it's not always clear whether a person's words are actually announcing a disaster.
- We will build a ML model to predicts which Tweets are about real disasters and which one's aren't. It can be easily seen that this is a Binary Classification task along with some concepts of NLP (Natural Language Processing).



Binary Classification

Binary classification is the task of classifying the elements of a set into two groups on the basis of a classification rule. Logistic Regression is used for binary classification.



$$\hat{y}^{(i)} = \sigma(w^T x^{(i)} + b), \text{ where } \sigma(z^{(i)}) = \frac{1}{1 + e^{-z^{(i)}}}$$

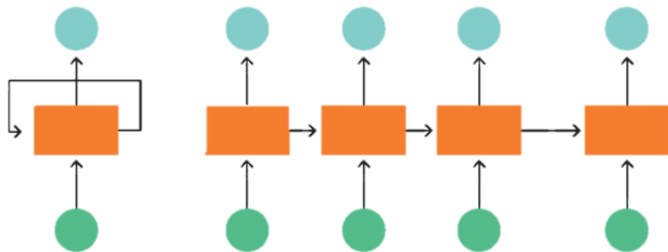
$$J(w, b) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^m [(y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}))]$$

Sequential Models

Sequential models are the machine learning models that input or output sequences of data. Sequential data includes text streams, audio clips, video clips, time-series data and etc.

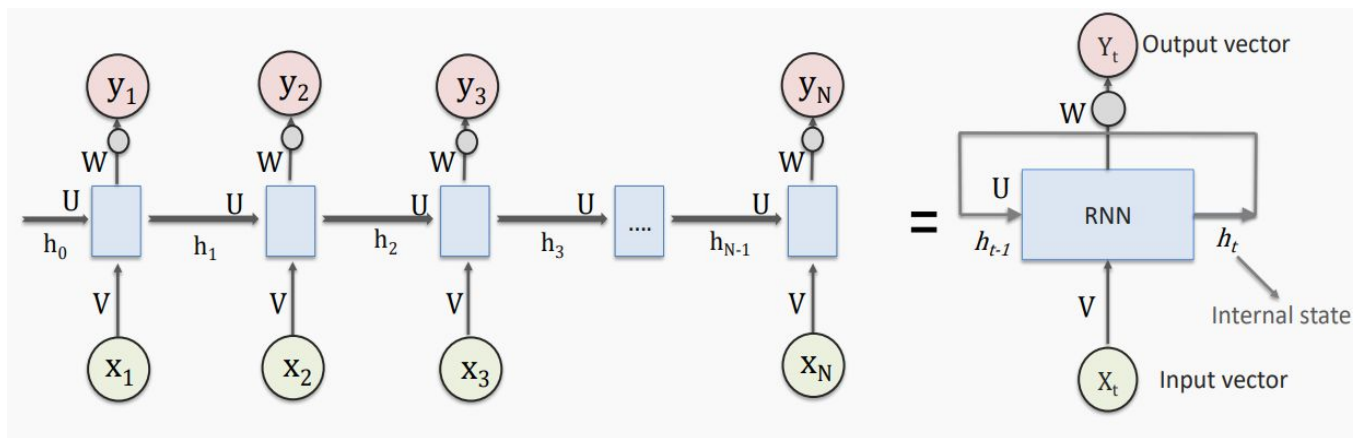
We can implement such sequential models using :-

- RNNs
- GRUs
- LSTMs

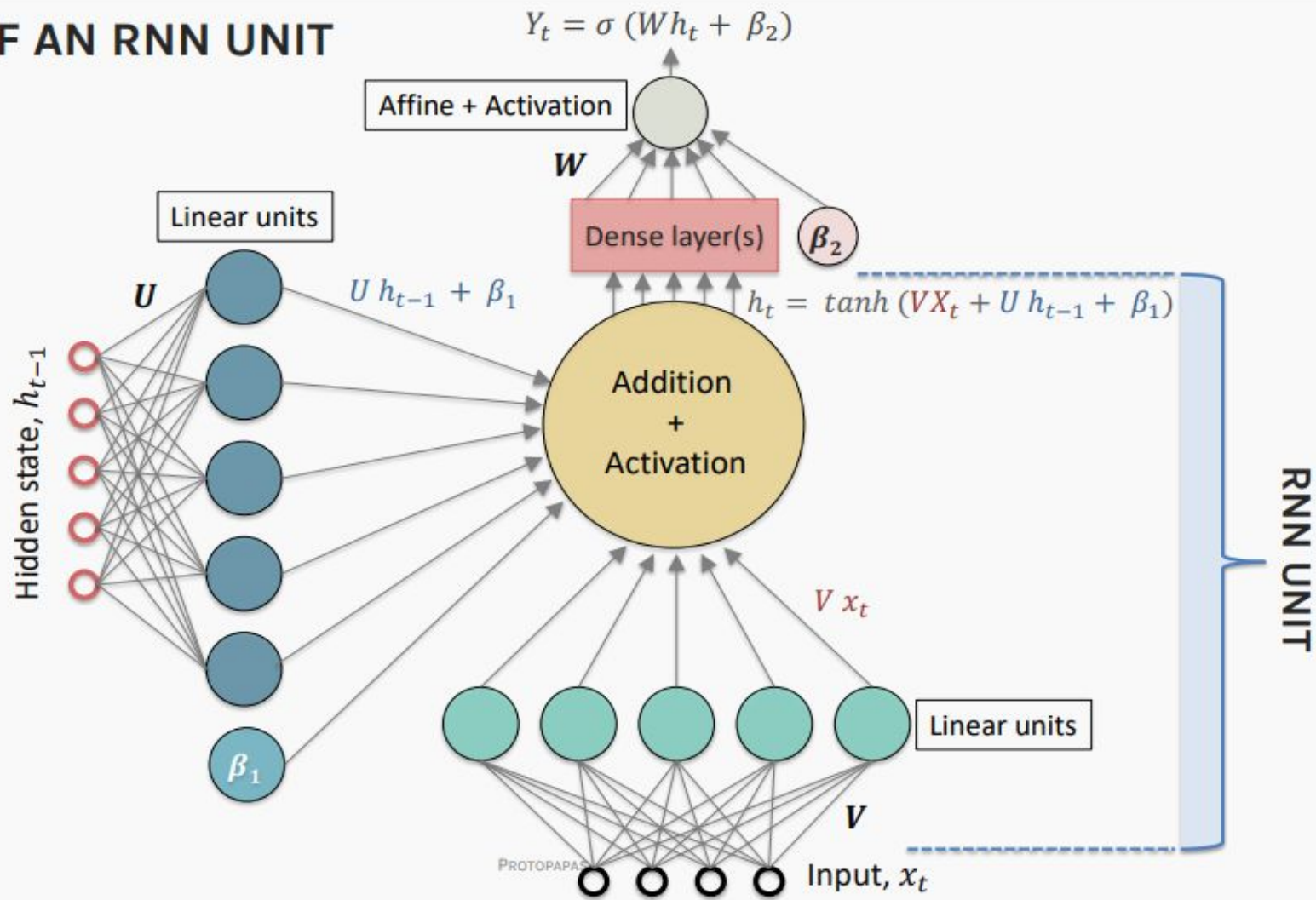


Recurrent Neural Network (RNN)

RNNs are governed by a recurrence relation applied at every time step for a given sequence.

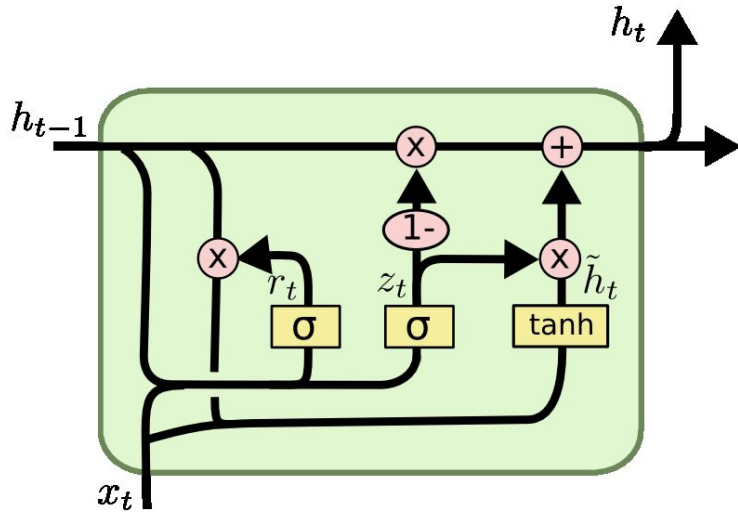


ANATOMY OF AN RNN UNIT



Gated Recurrent Unit (GRU)

A gated neural network uses processes known called update gate and reset gate. This allows the neural network to carry information forward across multiple units by storing values in memory. When a critical point is reached, the stored values are used to update the current state.



$$\mathbf{R}_t = \sigma(\mathbf{V}_R \mathbf{X}_t + \mathbf{U}_R \mathbf{h}_{t-1} + \beta_R)$$

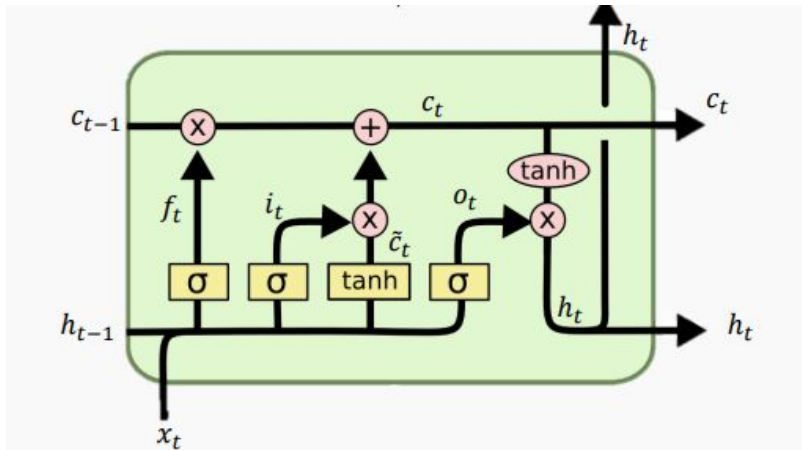
$$\mathbf{Z}_t = \sigma(\mathbf{V}_Z \mathbf{X}_t + \mathbf{U}_Z \mathbf{h}_{t-1} + \beta_Z)$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{V} \mathbf{X}_t + \mathbf{U} [\mathbf{R}_t \odot \mathbf{h}_{t-1}] + \beta_1)$$

$$\mathbf{h}_t = \mathbf{Z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\mathbf{h}}_t$$

Long Short Term Memory (LSTM)

LSTM is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. It uses three gates - input gate, output gate and forget gate.



$$\begin{aligned} f_t &= \sigma(V_f X_t + U_f h_{t-1} + \beta_f) & c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ i_t &= \sigma(V_i X_t + U_i h_{t-1} + \beta_i) & o_t &= \sigma(V_o X_t + U_o h_{t-1} + \beta_o) \\ \tilde{c}_t &= \tanh(VX_t + Uh_{t-1} + \beta_1) & h_t &= o_t \odot \tanh(c_t) \end{aligned}$$

TOKENIZER

Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords.

Given a sentence or paragraph, space tokenizer tokenizes into words by splitting the input whenever a white space is encountered.

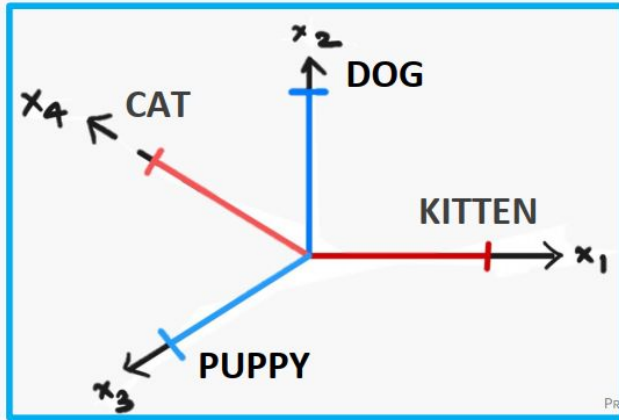
The tokenizer in Tensorflow tokenizes the sentence as well as assigns a unique number to each token.

An example of a sentence being tokenized and padded with zeros is:

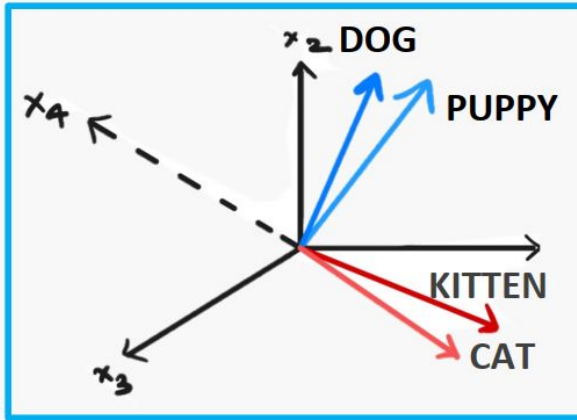
```
Original Sentence - Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all
The tokenized sequence - [ 120 4634 25 5 869 9 22 264 139 1620 4635 90 41 0
0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0]
```

Word Embeddings

Word embedding is a representation of a word. The embedding is used in text analysis. Typically, the representation is a real-valued vector that encodes the meaning of the word in such a way that words that are closer in the vector space are expected to be similar in meaning.

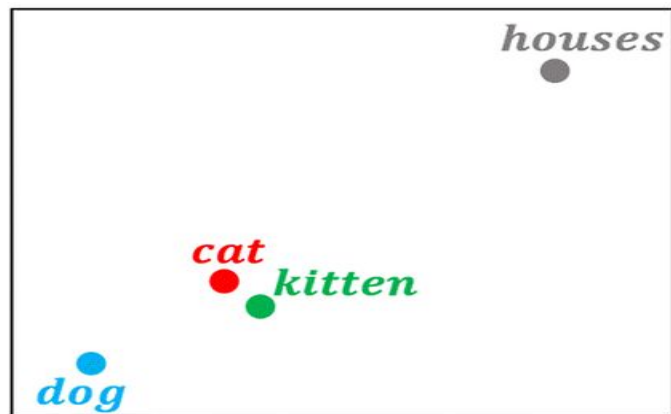


PROTOPAPAS



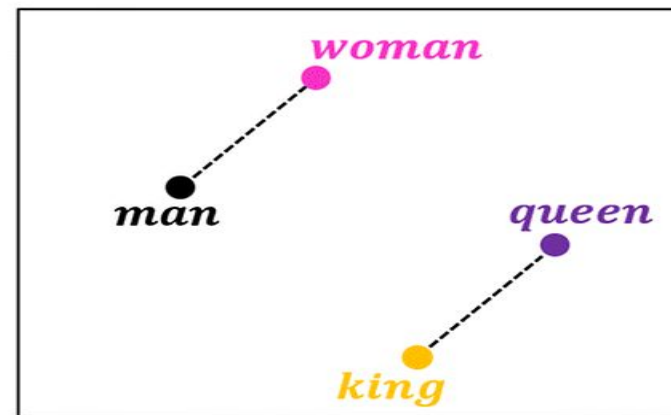
	living being	feline	human	gender	royalty	verb	plural
<i>cat</i> →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
<i>kitten</i> →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
<i>dog</i> →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
<i>houses</i> →	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8

Dimensionality
reduction of
word
embeddings
from 7D to 2D



<i>man</i> →	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
<i>woman</i> →	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
<i>king</i> →	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
<i>queen</i> →	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

Dimensionality
reduction of
word
embeddings
from 7D to 2D



Word

Word embedding

Dimensionality
reduction

Visualization of word
embeddings in 2D

Exploratory Data Analysis

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1
...
7608	10869	NaN	NaN	Two giant cranes holding a bridge collapse int...	1
7609	10870	NaN	NaN	@aria_ahrary @TheTawniest The out of control w...	1
7610	10871	NaN	NaN	M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt...	1
7611	10872	NaN	NaN	Police investigating after an e-bike collided ...	1
7612	10873	NaN	NaN	The Latest: More Homes Razed by Northern Calif...	1

7613 rows × 5 columns

Some of the Disaster Tweets are -

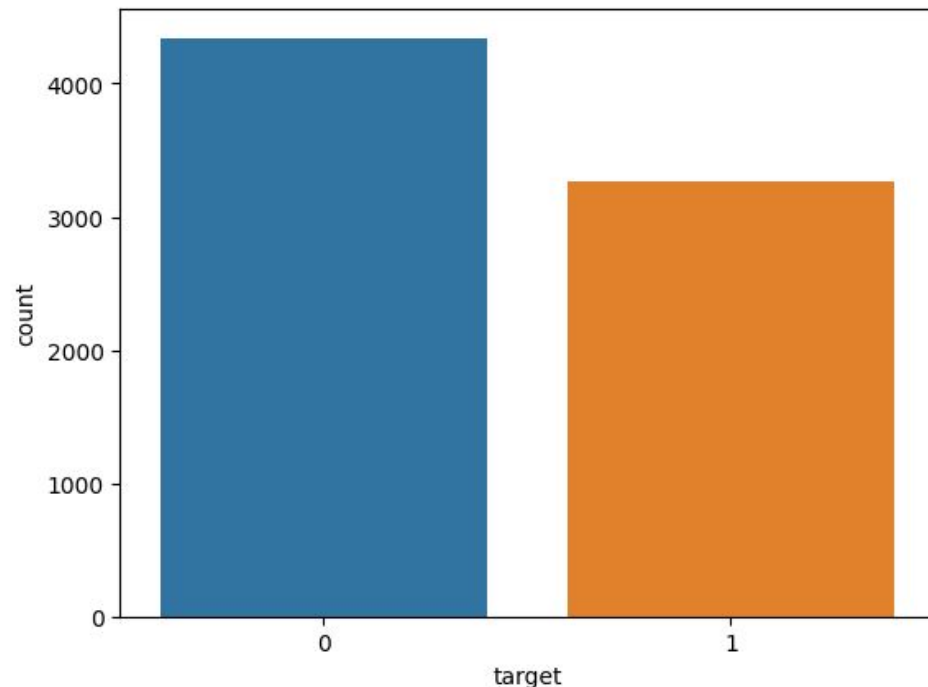
Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all

Forest fire near La Ronge Sask. Canada

All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected

13,000 people receive #wildfires evacuation orders in California

Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school



Some of the Non - Disaster Tweets are -

What's up man?

I love fruits

Summer is lovely

My car is so fast

What a goooooooooaaaaaaal!!!!!!

Creating Training Data

```
# Some important initializations
vocab_size = 31925      # Number of unique words + 1
max_len = 31           # Max length of the sentences
oov_token = "<OOV>"     # Represents the Unknown words
embedding_dim = 30
```

```
# Creating a tokenizer
tokenizer = Tokenizer(num_words=vocab_size, oov_token=oov_token)
tokenizer.fit_on_texts(training_sentences)
word_index = tokenizer.word_index
sequences = tokenizer.texts_to_sequences(training_sentences)
padded_sequences = pad_sequences(sequences, padding='post', maxlen=max_len)
```

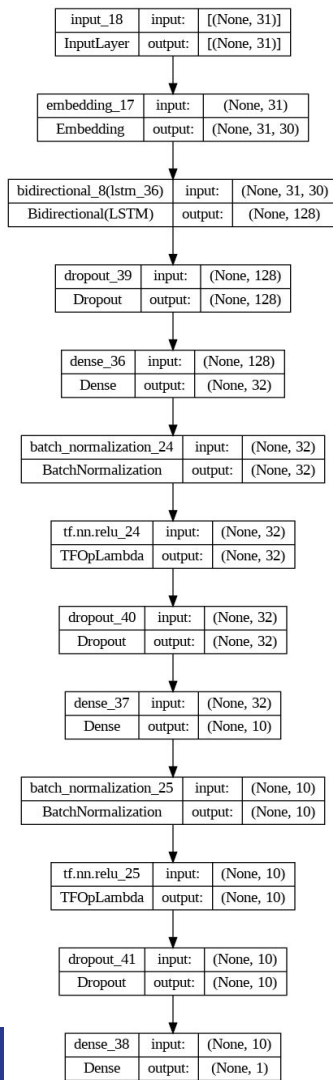


```
print("Original Sentence - ", df['text'][0])
print("The tokenized sequence - ", padded_sequences[0])
```

```
Original Sentence - Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all
The tokenized sequence - [ 120 4634 25 5 869 9 22 264 139 1620 4635 90 41 0
0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0]
```

```
print("Original Sentence - ", df['text'][1])
print("The tokenized sequence - ", padded_sequences[1])
```

```
Original Sentence - Forest fire near La Ronge Sask. Canada
The tokenized sequence - [ 190 46 230 800 6955 6956 1405 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0]
```

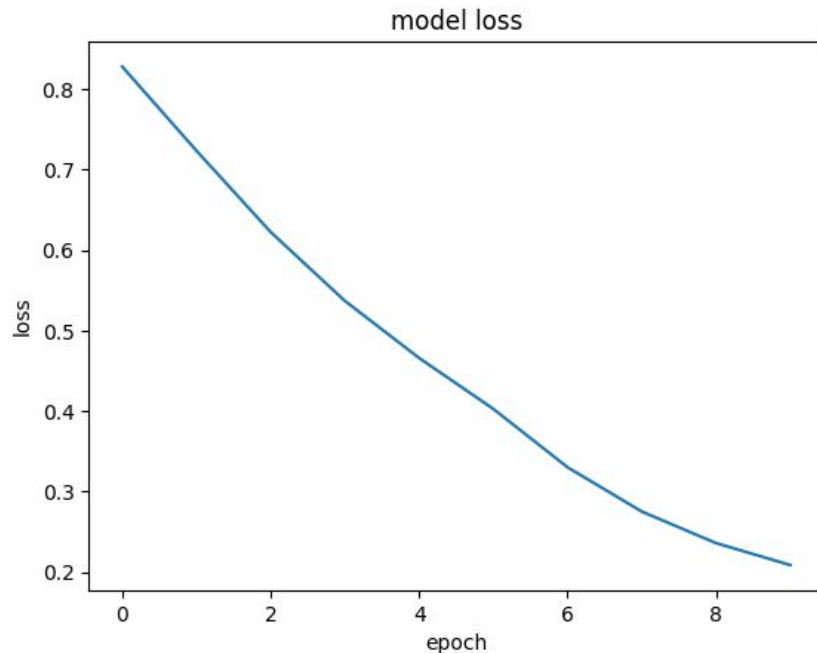
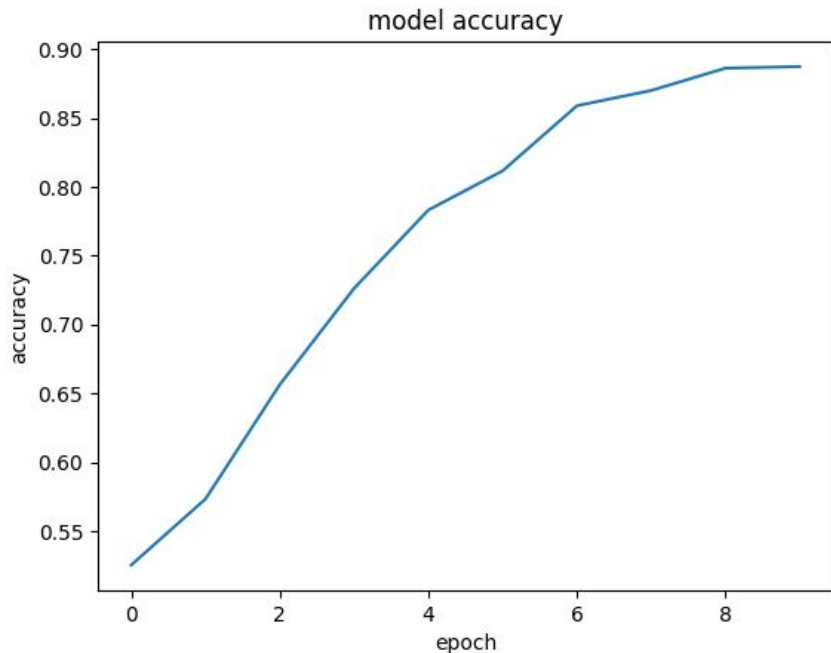



Creating the Model

Results on Training Data

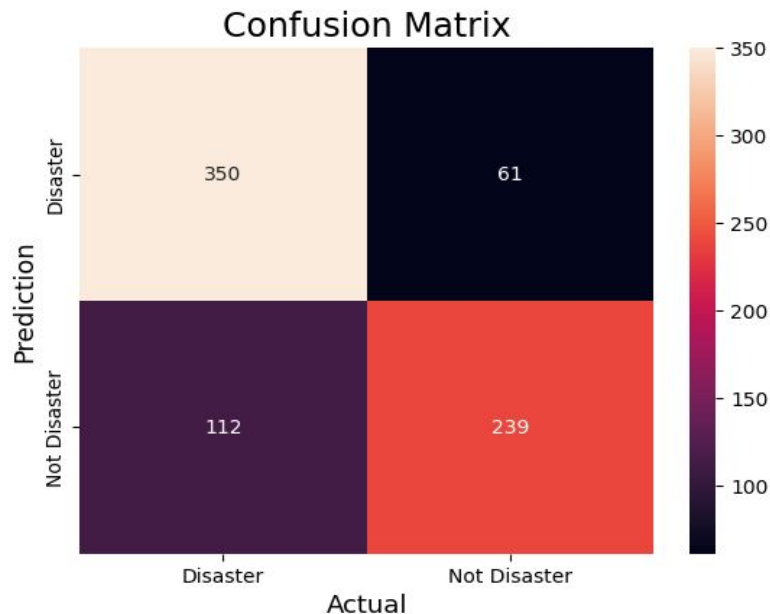
Epoch 10/10

54/54 [=====] - 2s 40ms/step - loss: 0.2135 - accuracy: 0.9583



Results on Testing Data

24/24 [=====] - 1s 4ms/step - loss: 0.5821 - accuracy: 0.7730
[0.5821198225021362, 0.7729659080505371]



Recall = 0.6809116809116809
Precision = 0.7966666666666666
F1 Score = 0.7342549923195083

Conclusion

- We studied the concepts of regression and binary classification. Then we studied some concepts of Natural Language Processing including RNNs, GRUs and LSTMs, word embeddings.
- We trained a sequential model using Tensorflow on the training data and achieved a high accuracy of around 95%.
- On the test data also, we were able to get a good accuracy of around 78% which is much higher than benchmark accuracy.



Improvements and Future Scope

- We can use different tokenizers and some pre-trained tokenizers which can definitely improve the results.
- We can clearly see that the model is undergoing overfitting, so we can use some techniques like early stopping, remove some layers, some regularization methods etc. to remove overfitting.
- We can use pretrained word embeddings like GLove, Word2Vec, etc.
- Some state of the art models like BERT transformer can be used.



Thank You

