Comments Toxicity Detection

Dr. Baishali Dutta

Outline

Motivation

Life in 21st Century

Motivation

Negative Effects of Social Media

More children 'self-harming because of cyber-bullying'

Cyberbullying knows no age limit

Missouri woman indicted in case involving MySpace-related suicide

CYBER BULLYING AFFECTS REAL LIVES !

New cyber-bullying weapon: Mobile phones

Widespread use of handsets make youths an easy target

THE CCGS NEWS

'CYBERBULLYING RUINES LIFES'

Death by social media

South Africa has a cyberbullying rate of 24% which places it at roundabout four in the world

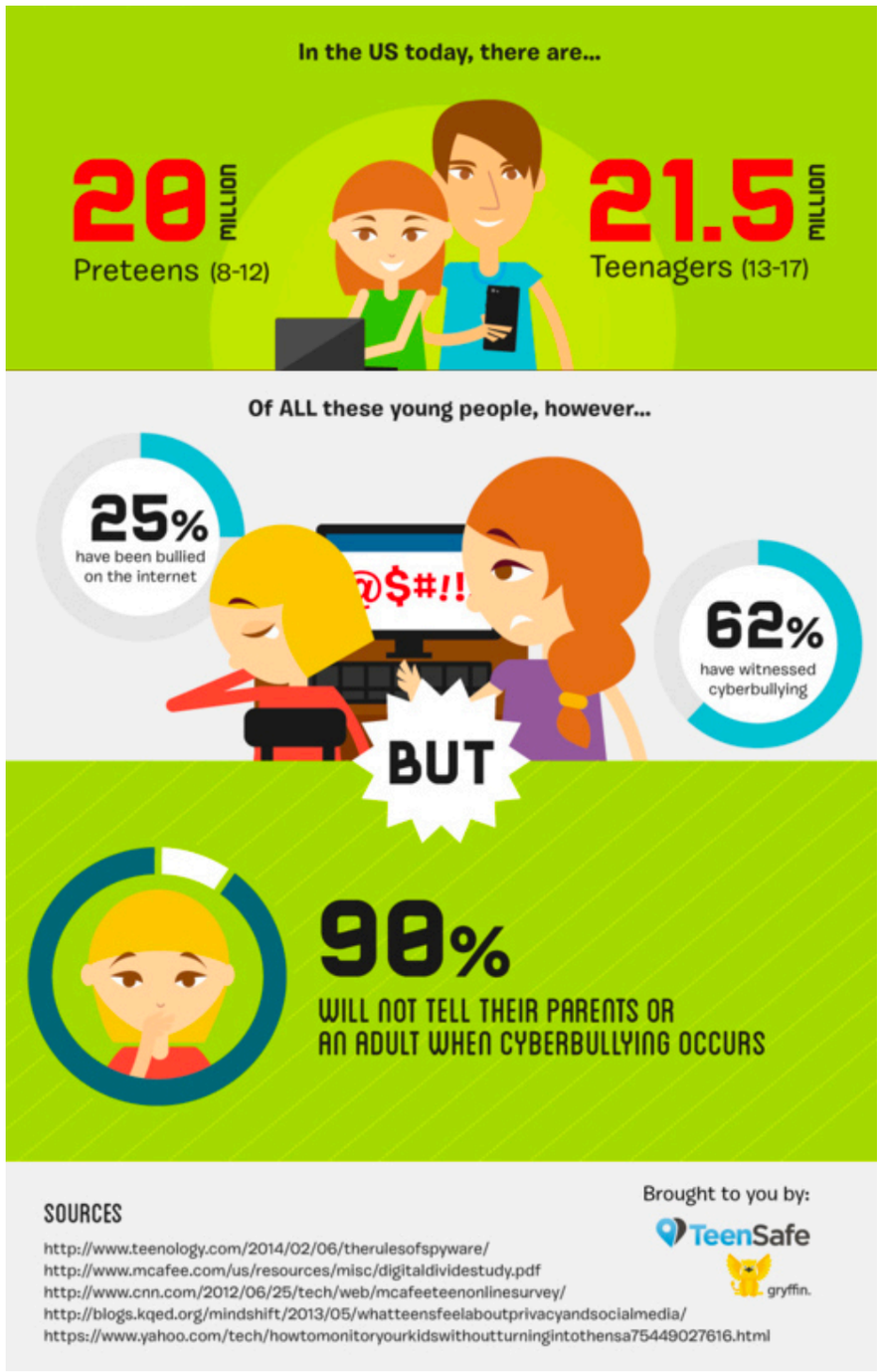
TIMES TRENDS

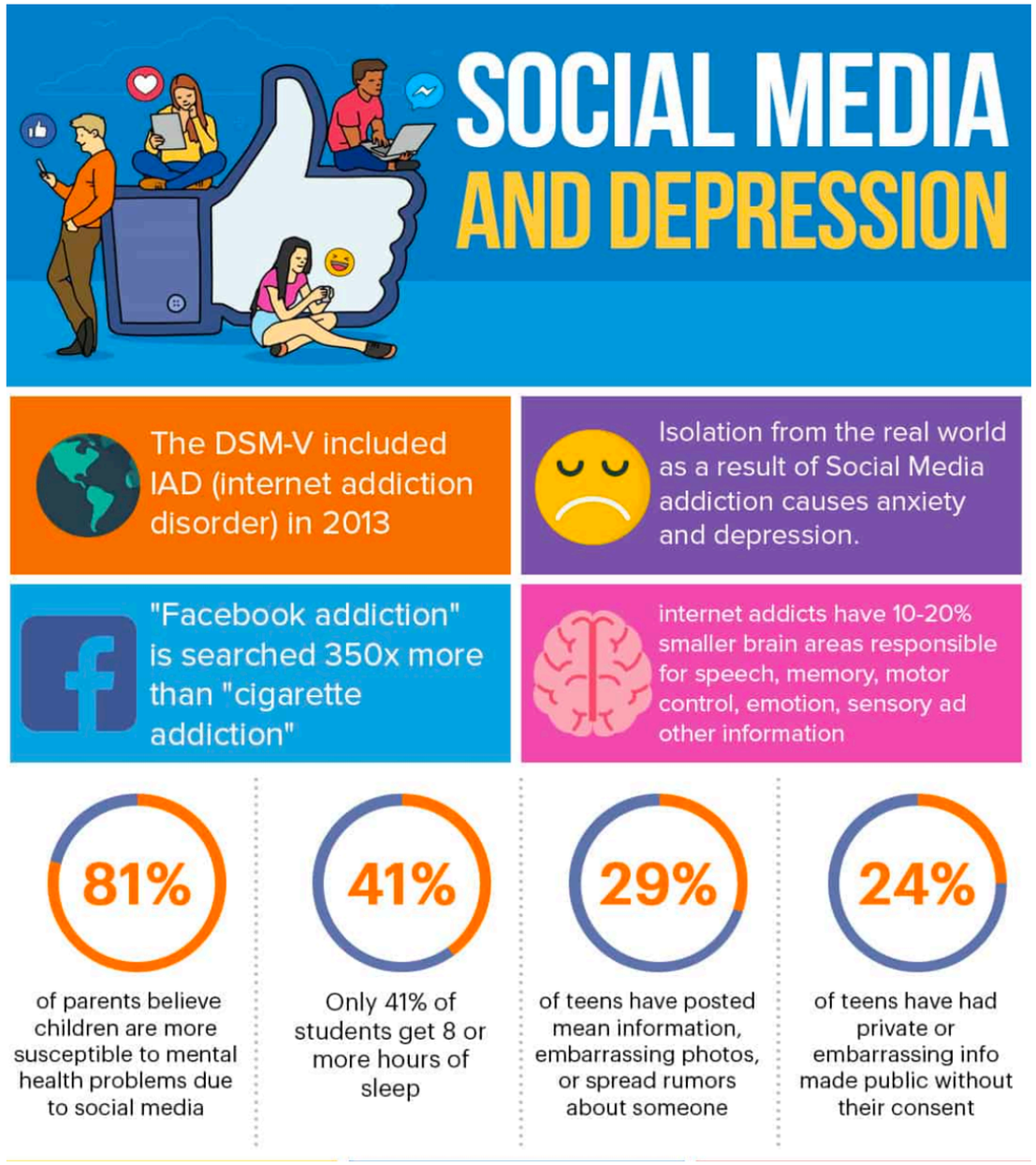
Online abuse becoming part of kids' life

4 In 10 Parents Unable To Help; Schools Urged To Teach How To Tackle Cyberbullying

Motivation

Surveys





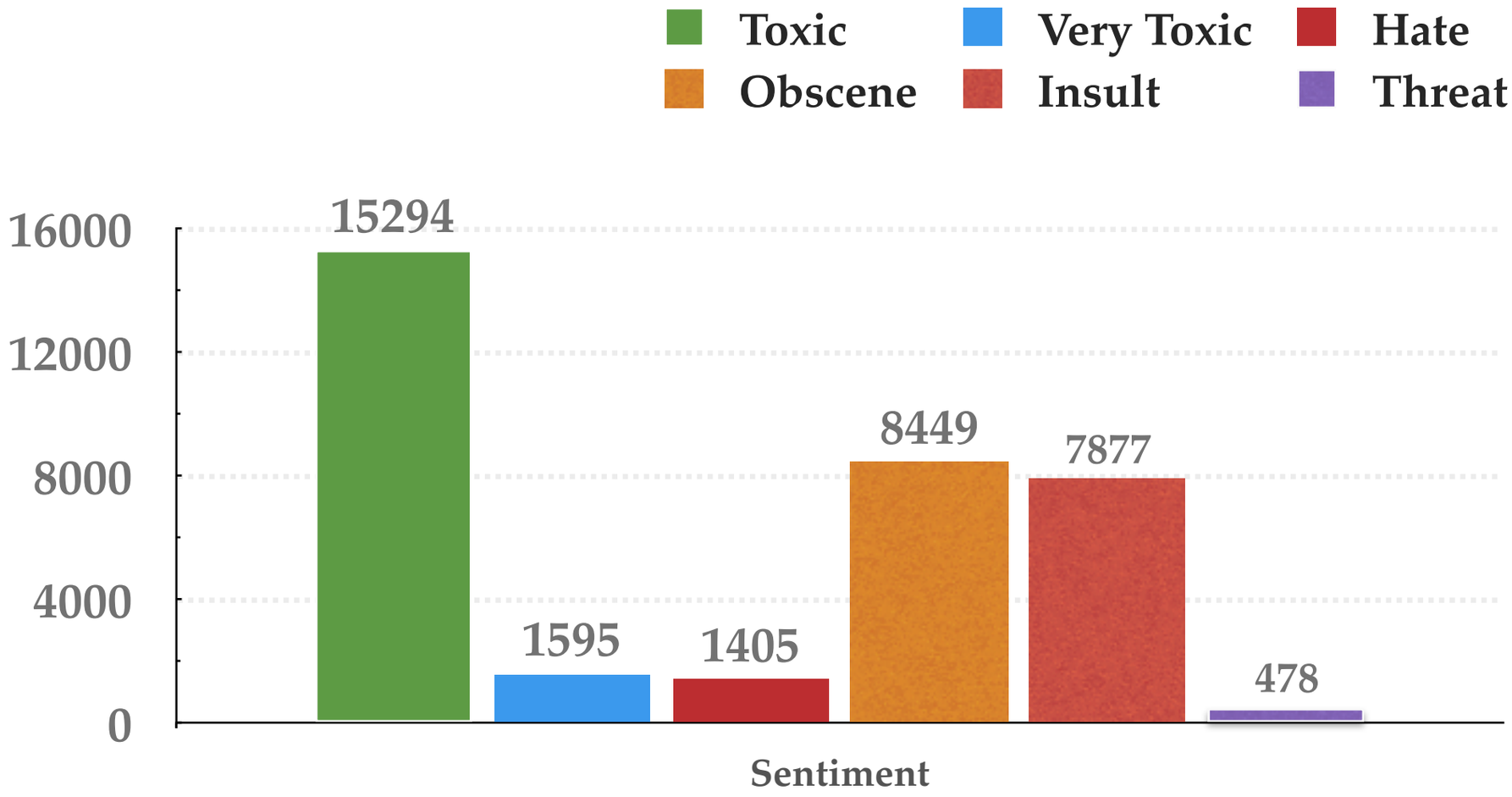
Motivation

Prevent Cyberbullying
Reduce Toxicity

Motivation

Preventive Measures

Exploratory Data Analysis (EDA)



- Training Data: Total: 159571 entries, Neutral: 124473 (none of the toxicity labels assigned)
- Huge unbalanced dataset, no null entries and empty strings present
- Multi-Label Classification Problem, target labels are not mutually exclusive, i.e. More than one right answer

Data Preprocessing

Text Processing

Lower Case All Words

- lower, camel or upper case written words treated as a same word
- all, All, ALL converted to all

Removed Punctuations

- . , ; : etc.
- Do not add any helpful information

Contraction Mapping

- **Contracted words are expanded**
- **Example : aren't : are not, can't: cannot etc.**

Removed Emojis

- Removed emoticons, symbols, flags etc.
- Could be helpful for certain cases of sentiment analysis

Fixed Misspelled Words

- Corrected with TextBlob library

Removed Stopwords

- *the, in, among, for, where* etc.
- Do not add any helpful information

Lemmatisation

- Finds the base form of words (spacy)
- Debatable in Sentiment Analysis as it can reaks parts-of-speech tagging and alters polarity of a word
- Reduces the word corpus

Data Preprocessing

Embedding Vectors

- Pre-trained word embedding vectors Glove.6B is used

Tokenizer

- Tokenizer (from keras) is used to tokenize the text with max vocab size of 20000
- Each word is assigned to the corresponding feature vector from the Glove.6B

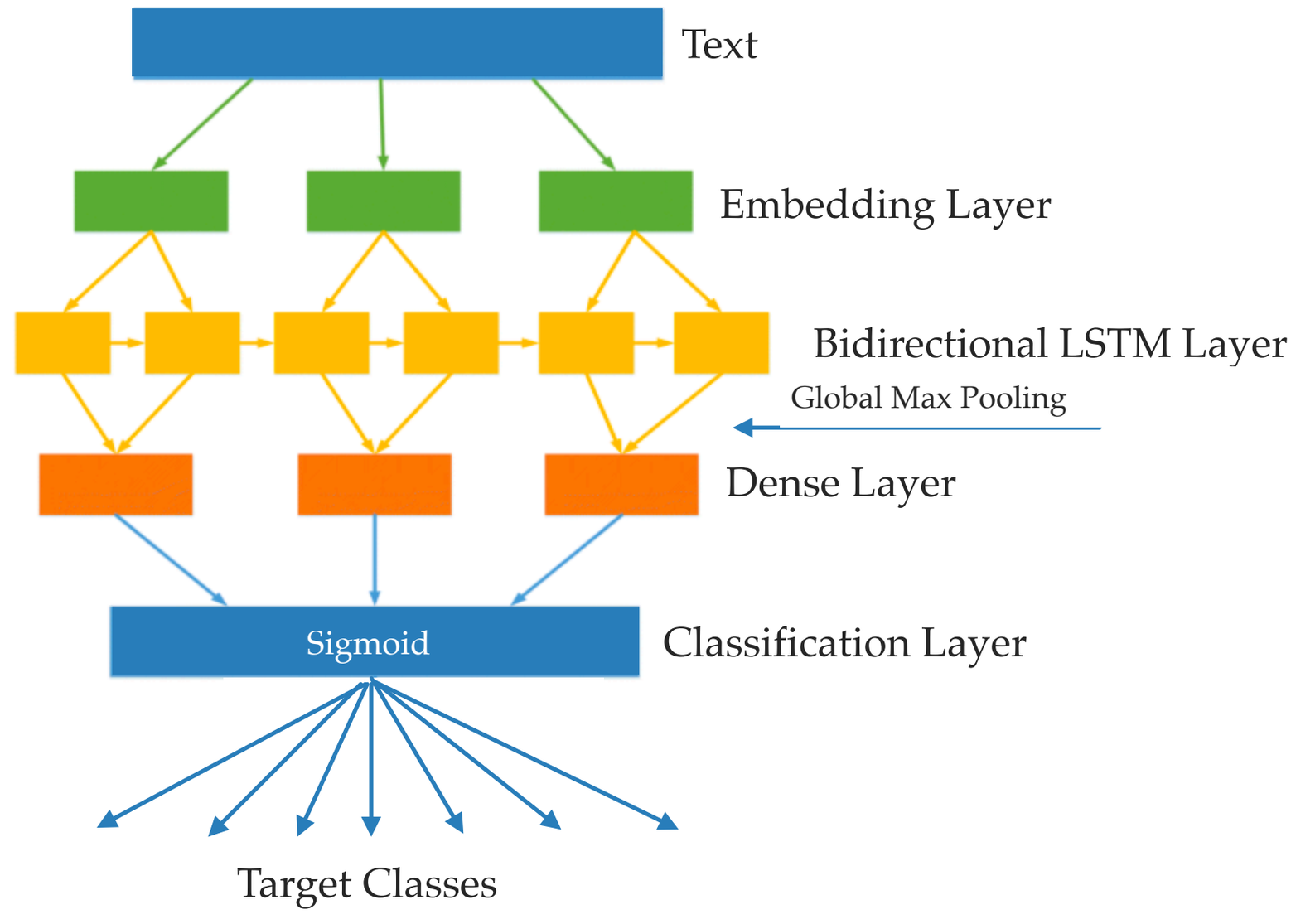
Deep Learning Model (Bi-directional LSTM)

- Recurrent Neural Network (RNN) is used as it was primarily built to tackle NLP problems (sequential data, sentences are sequence of words)

- In particular, Bidirectional Long Short-Term Memory (LSTM) RNN model is chosen

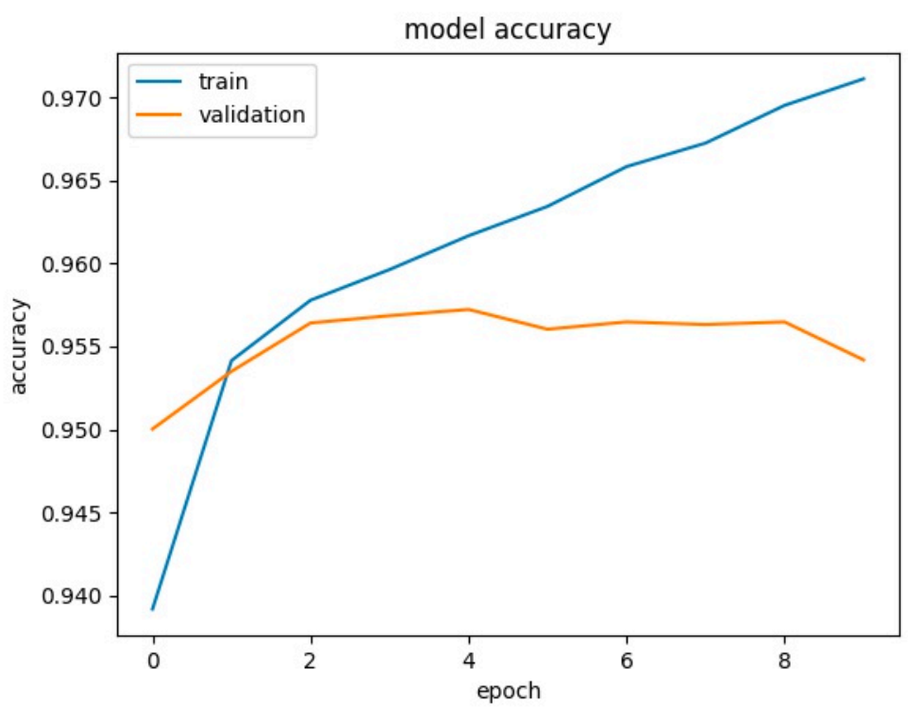
- LSTM can remember longer sequences than regular RNN

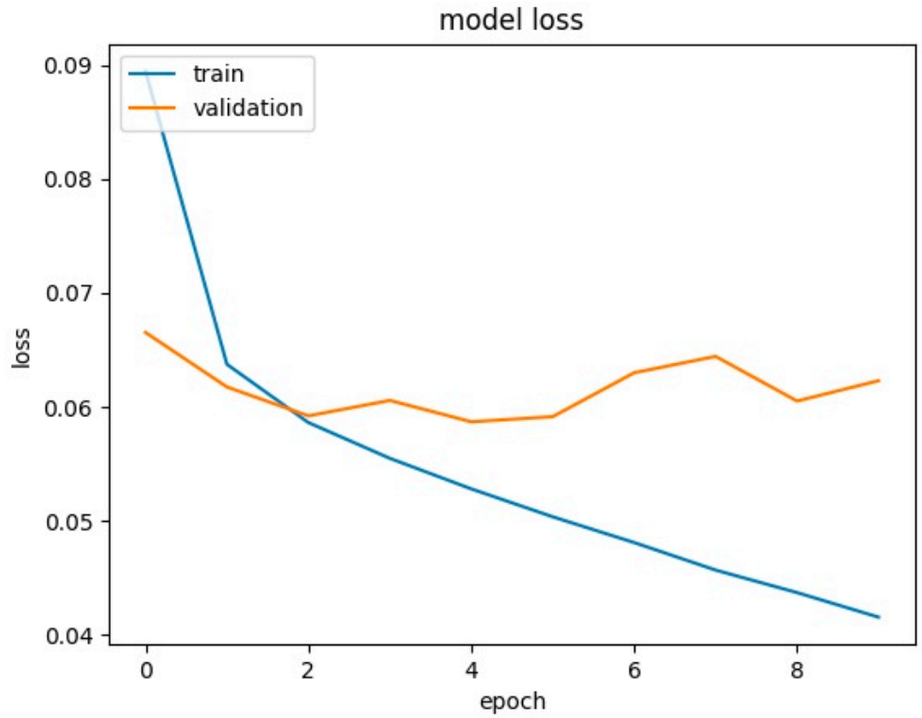
- Making it bidirectional, helps in a way that it can see at a given sequence both previous and next sequences in the text/ sentence



Model Evaluation

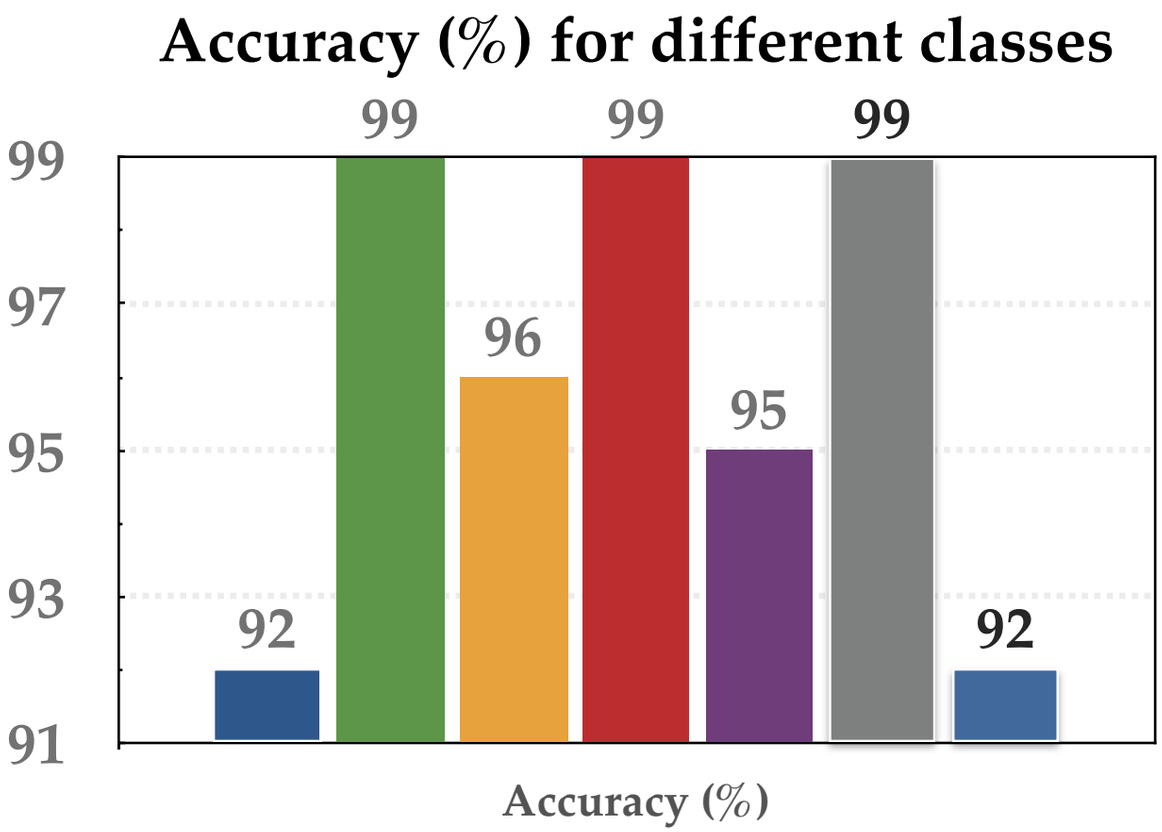
- 20% of training dataset is kept for validation and rest used for training
- Accuracies on both training and validation sets and loss in each epoch (used early stopping monitoring validation loss)

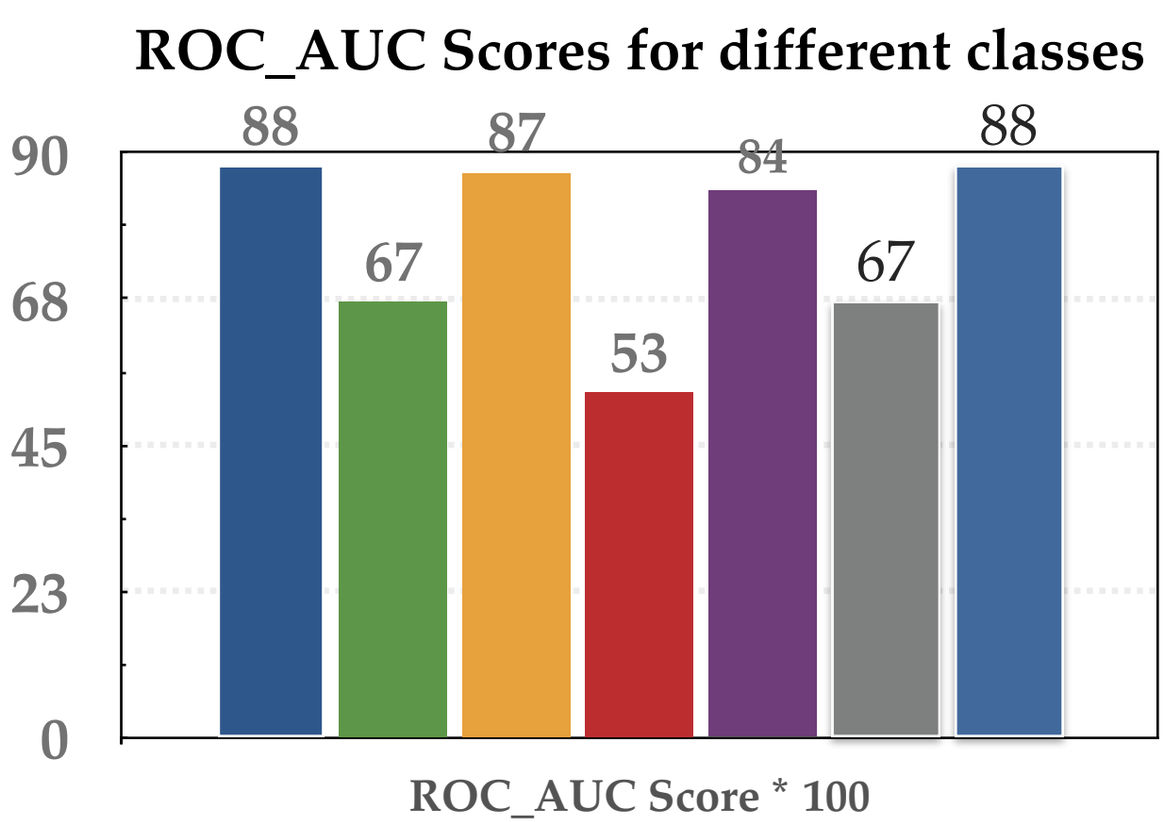




- Epoch = 4 is chosen as the best validation loss, patience = 5 indicates another 5 epochs that were run to confirm if the validation loss improves
- Next step is to evaluate the model on the test data (499 data points)

Model Evaluation





Mean Accuracy: 96.3%

Mean ROC_AUC Score: 0.76

DEMO