CS3244-04

Toxic Comments Classification

With the evolution of technology, platforms, such as social media,that allows the communication of personal thoughts and feelings are increasingly prevalent. However, this degree of freedom is associated with problems such as promoting hate, hurling abuse anonymously or cyber-bullying - resulting in a toxic online community.

Hence, this project aims to come up with a multi-headed model to distinguish toxic comments on Wikipedia from clean ones, and to identify the types of toxicity present.

Proposed Model

We propose a neural network which consists of an input layer, an embedded layer, a LSTM (Long Short Term Memory) layer, a normal layer with ReLU as activation function and an output layer with sigmoid as the activation function, using python's keras library.

LSTM was employed as the model preserves the semantic meaning of the training data, which we feel is an important factor when determining if a comment is toxic or not. As opposed to regular RNNs, they are able to learn long-term dependencies.

Preprocessing

Keras Tokenizer

- *Turn all words to lowercase and to filter all special characters out*
- *Each training example is then turned into a sequence of integers, where each integer is unique to each word in the training example.*
- *Padded all the training examples to the same length to ensure that all the training examples have equal dimensions. (Requirement for LSTM layer)*

Training and validation

- *we used 90% of the training data to perform training of the model, and the remaining 10% for validation.*
- *To reduce overfitting, we employed :*
  - *Dropout , which randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much, reducing overfitting.*
  - *Global Max Pooling, minimize overfitting by reducing the total number of parameters in the model by reducing the spatial dimensions of a three-dimensional tensor.*
- *As a measure of loss, we use binary Cross-Entropy, as our network's outputs can be thought of as representing independent hypotheses (classification on toxic, obscene, insult, etc) and the node activations (Sigmoid) can be understood as representing the probability that each hypothesis might be true.*

Results

Metric used in this competition : AUROC (Area Under Receiver Operating Characteristic Curve)

- *A performance measurement for classification problems at various thresholds settings. It tells us how much the model is capable of distinguishing between classes.*

96.87%

With our model, we have obtained a score of 0.9687 on Kaggle.

Evaluation

While our model obtained a satisfactory score , it is by no means the best model.

Looking at the confusion matrix on the right , you will notice that our model fails to classify any comments that are threats. It did quite badly on severe toxic and identity hate as well.

As clean comments accounts for about 90% of our training data, our model can score decently due to the fact that it can classify non-toxic comments well and not toxic comments.

Even though it did well on this Kaggle Competition, this is obviously not suitable for production as the consequences of wrongly classifying toxic comments as clean is catastrophic.

Done on a 80-20 split of our training data

Column: Toxic

Predicted	0	1	All
True			
0	28267	592	28859
1	493	2563	3056
All	28760	3155	31915

Column: Threat

Predicted	0	1	All
True			
0	31841	0	31841
1	74	0	74
All	31915	0	31915

Column: Severe Toxic

Predicted	0	1	All
True			
0	31545	49	31594
1	233	88	321
All	31778	137	31915

Column: Insult

Predicted	0	1	All
True			
0	29948	353	30301
1	404	1210	1614
All	30352	1563	31915

Column: Obscene

Predicted	0	1	All
True			
0	29962	238	30200
1	312	1403	1715
All	30274	1641	31915

Column: Identity hate

Predicted	0	1	All
True			
0	31615	6	31621
1	272	22	294
All	31887	28	31915

Sponsered by

SGINNOVATE · INFOCOMM MEDIA DEVELOPMENT AUTHORITY · sea connecting the dots · indeed

Project by:

LIM WEI LIANG BARRY (SCI2)
LEE YI QUAN (SCI3,COM3)
LEE JUN HAN (BRYAN) (CEG3)
MUHAMMAD AFIF B MOHD ALI (SCI2)
ISABEL ANG YUET TING (SCI2)
TAN CHEE WEE (COM3)