Deep learning project:
Multi-label Toxic comment classification

Vishwas Prabhu
Shubham Thakur

Identify and classify toxic online comments

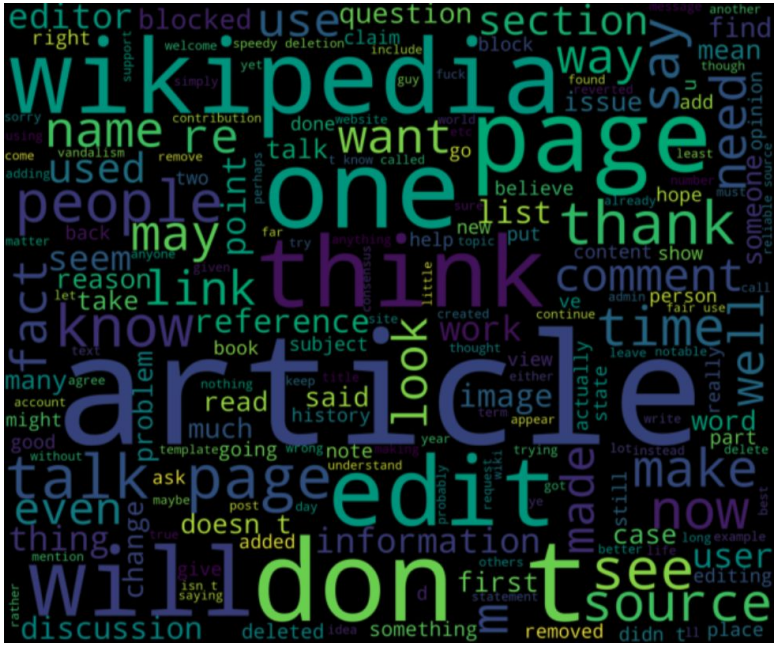
Motivation: Toxic comments on a platform can be defined as user comments that are rude, disrespectful or something that disengages other users from using the platform. Toxic, abusive and online harassment is an issue that plagues online platforms and results in users to stop using the platform leading to business loss and drop in daily active users. Platforms thus needs to identify such user comments and take action on time to avoid negative fallout and user churn from the platform.

Challenge: To build a multi-class NLP classifier which identifies different types of toxic comments such as threats, obscenity, insults, identity-based hate etc since some platforms are fine with certain types of profanity but not other type of toxic content/comment.

Data: Wikipedia comments which have been labeled by human raters for toxic behavior into Toxic, severe_toxic, obscene, threat, Insult, identity_hate.

https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

References:



JIGSAW

Data Cleaning

Common methods we used to clean the data

- Removed Website Links -> `r'https?://\S+|www\.\S+'`
- Removed HTML tags -> `r'<[^>]*>'`
- Removed Non ASCII characters -> `r'[^\x00-\x7E]+'`
- Removed Special Characters-> `r'["#$%&\'()\*\+-/:;<=>@\[\]\\\^_`{|}~]'`
- Removed Stop words
- Lowercase the sentence
- Removed Numbers -> `'\d+'`

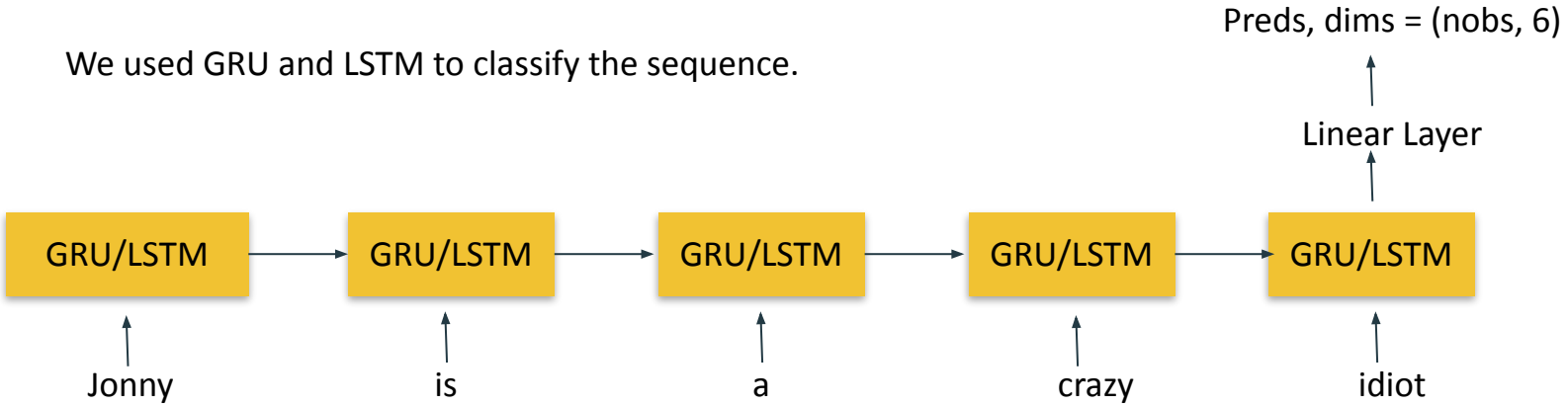
Example:

```
text1 = "I had such... high hopes for this dress!!!!"

Output = 'I had such high hopes for this dress'
```

Model Building

We used GRU and LSTM to classify the sequence.



Model Configurations

lr = 0.01
Batch Size = 1000
lossFun = nn.BCEWithLogitsLoss()
Optimizer = Adam
GRU/LSTM hidden layer = 100

Labels	toxic	severe_toxic	obscene	threat	insult	identity_hate
Actual	1	1	0	1	0	0
Pred	0.9	0.8	0.02	0.4	0.2	0.1

NLP pipelines and evaluation metrics:

Metrics

Model	Training - AUC	Test - AUC
Random forest (CBOW)	0.55	0.55
GRU	0.98	0.97
LSTM	0.97	0.96

Possible Improvements

- Current model uses pretrained embedding. The model embedding could be fine tuned while training.
- BERT pretrained model for classification.
- Learning rate Scheduler
- Fine Tune parameters like LR, Linear layer size etc.

Identify and classify toxic online comments

Motivation:

- Toxic comments on a platform can be defined as user comments that are rude, disrespectful or something that disengages other users from using the platform. Toxic, abusive and online harassment is an issue that plagues online platforms and results in users to stop using the platform leading to business loss and drop in daily active users.
- Platforms thus needs to identify such user comments and take action on time to avoid negative fallout and user churn from the platform.

Challenge:

- To build a multi-class NLP classifier which identifies different types of toxic comments such as threats, obscenity, insults, identity-based hate etc since some platforms are fine with certain types of profanity but not other type of toxic content/comment.

Data: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

- Wikipedia comments which have been labeled by human raters for toxic behavior into Toxic, severe_toxic, obscene, threat, Insult, identity_hate.

References:

JIGSAW

Conversation AI

Data Description and NLP pipelines:

Size of the Dataset

Train Data: 159,571

Test Data: 153,164

Distribution of Labels

toxic	severe_toxic	obscene	threat	insult	identity_hate	Non Toxic
15,294	1,595	8,449	478	7,877	1,405	124,473

Data Processing

- Tokenization
- Remove Stop Word
- Stemming/lemmatization

Feature Extraction

- Bag of Word
- TF-IDF
- Word2Vec

Data Modelling

- ML models like dtree, rf
- RNN
- LSTM

Deep learning project proposal

Vishwas Prabhu
Shubham Thakur

Thank You