Bots and Gender Profiling with Convolutional Hierarchical Recurrent Neural Network

Juraj Petrik & Daniela Chuda

Our background

Source code plagiarism detection
String similarity and frequency analysis

Source code authorship attribution
Deep learning and frequency analysis

Preprocessing

Emoji translation from unicode to word description.

Lemmatization for better generalization in small datasets.

Tokenization to divide text to tokens, in our case we can talk about words.

Stop words removal, because in theory they got very little information value.

Tokens encoding and zero padding are necessary for embedding layer of fixed length.

Before:
@0rangelic @Roslinnovation You're doing way better than everyone here. 😉

After emoji translation and lemmatization:
@0rangelic @Roslinnovation -PRON- be do way well than everyone here . : winking_face :

Classifier

Embedding layer to create word vector represenation.

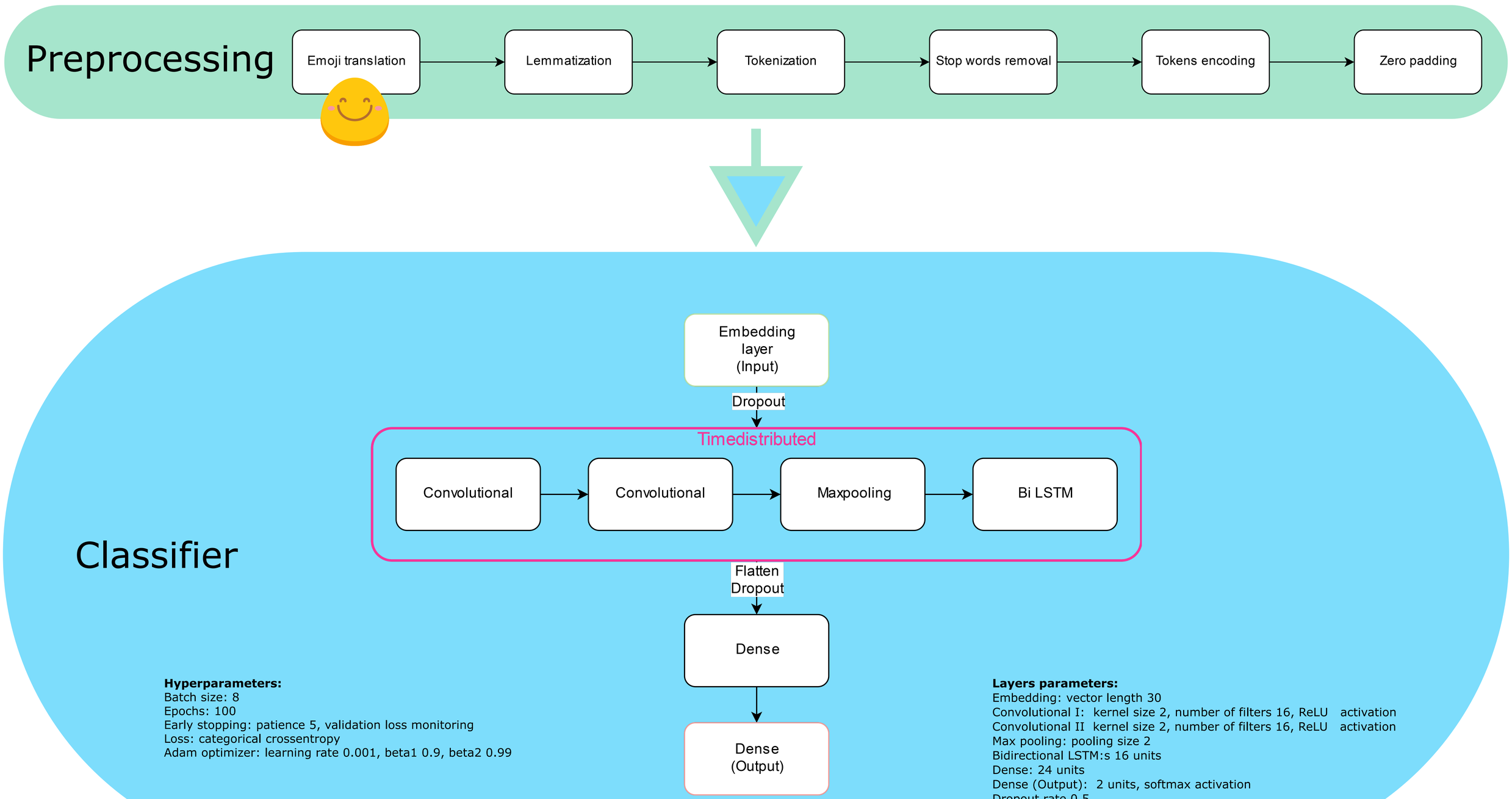
Convolutional layers act practically as feature extractors.

Maxpooling layers are used for dimensionality reduction.

Dropouts prevent overfitting by randomly dropping nodes connections.

Bidirectional LSTM processes tweets from right to left and from left to right. RNN networks shows great results in text classification problems.

Hierarchy in recurrent layers helps to capture relations in multiple tweets sequences.



Hyperparameters:
Batch size: 8
Epochs: 100
Early stopping: patience 5, validation loss monitoring
Loss: categorical crossentropy
Adam optimizer: learning rate 0.001, beta1 0.9, beta2 0.99

Layers parameters:
Embedding: vector length 30
Convolutional I: kernel size 2, number of filters 16, ReLU activation
Convolutional II kernel size 2, number of filters 16, ReLU activation
Max pooling: pooling size 2
Bidirectional LSTM:s 16 units
Dense: 24 units
Dense (Output): 2 units, softmax activation
Dropout rate 0.5

Possible improvements

Tokenization (URLs, handles)

Extend vocabulary

Pretrained word embeddings

Hypernyms

Character level embeddings

STU FIIT
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

juraj.petrik@stuba.sk