Bots and Gender Profiling with Convolutional Hierarchical Recurrent Neural Network

Juraj Petrik & Daniela Chuda

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.

Classifier

- Embedding layer to create word vector representation.
- 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.

Possible improvements

- Tokenization (URLs, handles)
- Hypernyms
- Extend vocabulary
- Character level embeddings
- Pretrained word embeddings