Identifying Political Spectrum in News Articles
Markus Zechner and Halldora Gudmundsdóttir

Introduction and Motivation
- The main goal of this project is to automatically classify news articles based on their political spectrum.
- The political spectrum ranges from liberal (left) to conservative (right) and the classification is performed on the text body of the articles.
- But what are the direct implications of ranking news articles based on the political spectrum?
  - Automatically identifying the political spectrum of an article, the recommendation engine for users could be improved.
  - A variety of articles could be offered to users to provide them with different angles of a story.

Data, Labels and Cleaning
- The dataset is a collection of news articles (deepnews.ai) that originate from different news organizations ranging from very liberal to very conservative.
- The articles are labeled liberal, conservative or neutral using the website www Politics.com. It is important to point out that articles are labeled based on publisher and no manual labeling is involved.
- The train/dev/test split set is 70/15/15.
- The text body of each article is cleaned by (1) removing special characters, (2) converting letters to lower case, (3) splitting each sentence into words.

Models
- Pre-training: GloVe 30 dimensional word embeddings are used for word representation.
- Baseline: For each article word embeddings are averaged to a single vector which is fed into a softmax activation function.
- Fully Connected Neural Network: For each article word embeddings are averaged to a single vector which is fed into a 3-layered network with 600 neurons (relu activation), followed by a softmax output layer.
- LSTM Recurrent Neural Network:
- Conv1D Convolutional Neural Network
- Conv1D-LSTM Combination

Hyperparameter Tuning
- Because of computational limitations, tuning the models is performed in a sequential manner: (1) learning rate, (2) batch size, (3) article length, (4) number of neurons/filters and (5) regularization/dropout.
- The most sensitive parameters are the learning rate and article length.
- An example of tuning the Conv1D model:

Results
- LSTM generally works well but long training time limited hyperparameter tuning.
- Most models experience overfitting – dropout was used as a regularization.
- The larger training set (Dataset 2) shows a better performance – less overfitting and higher accuracy – but had very long training times (>30hrs).
- The best performance was achieved by the Conv Net, the LSTM does not work well for long sequences.
- A hybrid model Conv1D-LSTM showed promise in reducing the number of parameters but retaining the capabilities of the LSTM

Error Analysis
- It is harder for our model to quantify Left/Right than Left/Center or Right/Center.
- The Journalists correctly classify most extremes (Left vs. Right) but have trouble identifying Left/Center and Right/Center.
- Assuming the Journalists provides the ground truth (current labels are based on publishers) - the upper bound for accuracy is 65%

Future Work
- Investigate limitations of current labeling – algorithm might be learning style of writing rather than actual bias – ultimately get human labeled articles.
- Continue the error analysis:
  - Can the misclassified examples be used to get new labels?
  - Let a journalist mark phrases that were critical for the classification.
- Use an Attention model to assess the phrases that were critical for the algorithm in the classification task, use additional features (quotes)

References