Text Classification: 20newsgroups

Vaggelis Lamprou

Natural Language Processing

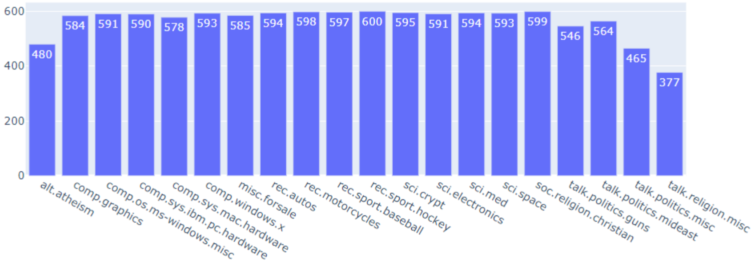
1. The Dataset: 20 newsgroups

- Training data: 11314 texts
  Test data: 7532 texts

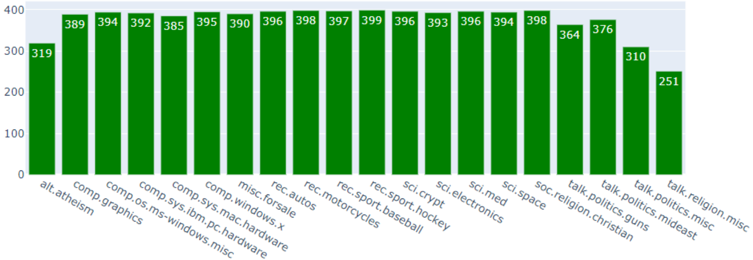
Categories :

- *alt.atheism*
- *comp.graphics*
- *comp.os.ms-windows.misc*
- *comp.sys.ibm.pc.hardware*
- *comp.sys.mac.hardware*
- *comp.windows.x*
- *misc.forsale*
- *rec.autos*
- *rec.motorcycles*
- *rec.sport.baseball*
- *rec.sport.hockey*
- *sci.crypt*
- *sci.electronics*
- *sci.med*
- *sci.space*
- *soc.religion.christian*
- *talk.politics.guns*
- *talk.politics.mideast*
- *talk.politics.misc*
- *talk.religion.misc*

Training set: Class distribution



Test set: Class distribution



2. ML approach

- Preprocessing
  - Preprocess1: lower characters, nltk's word_tokenize
  - Preprocess2: lower characters, nltk's word_tokenize, remove small words, remove stopwords, nltk's PorterStemmer
- Tfidf Vectorizer
  - tokenizer: Preprocess1, Preprocess2
  - n-grams: uni-grams, uni-grams & bi-grams, bi-grams
  - norm: 'l1', 'l2'
- Classifiers
  - Support Vector Machine
  - Multinomial Naive Bayes
  - Random Forest

- Evaluation of 36 pipeline models wrt validation accuracy score
- Optimal: SVM with Preprocess2, uni-grams & l2 norm
- Overall test scores
  - accuracy: 0.66
  - precision (weighted): 0.68
  - recall (weighted): 0.66
  - f1-score (weighted): 0.66

3. DL approach

- Kera's Tokenizer

  - Tokenize words and lower characters.

  - Learns 200-dim representations per text. Sequence of integers.

  - Consider 20K most common words and assign integers based on their frequency in descending order

- GloVe Embeddings

  - Pre-trained word vectors of dim 100

  - Matrix of shape (20K,100); eventually describes the weights of the Embedding layer of the NN

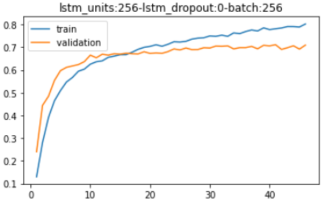
- The model: GloVe-based BiLSTM architecure

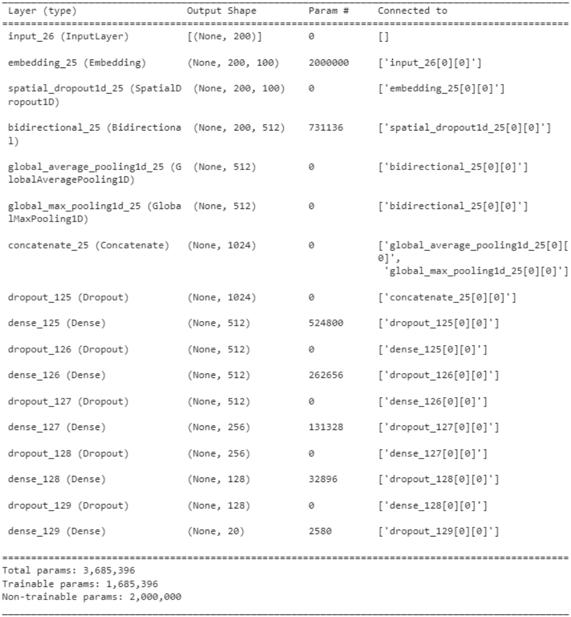
  - Hyper-params tuned: lstm nodes and dropout and training batch size

  - No need to tune: Adam(0.001), Categorical-Crossentropy Loss, Tanh activation fct

Accuracy history and Summary

- Overall test scores
  - accuracy: 0.67
  - precision (weighted): 0.68
  - recall (weighted): 0.67
  - f1-score (weighted): 0.67





```
__________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 input_26 (InputLayer)          [(None, 200)]        0           []

 embedding_25 (Embedding)       (None, 200, 100)     2000000     ['input_26[0][0]']

 spatial_dropout1d_25 (SpatialD  (None, 200, 100)    0           ['embedding_25[0][0]']
 ropout1D)

 bidirectional_25 (Bidirectiona  (None, 200, 512)    731136      ['spatial_dropout1d_25[0][0]']
 l)

 global_average_pooling1d_25 (G  (None, 512)         0           ['bidirectional_25[0][0]']
 lobalAveragePooling1D)

 global_max_pooling1d_25 (Globa  (None, 512)         0           ['bidirectional_25[0][0]']
 lMaxPooling1D)

 concatenate_25 (Concatenate)   (None, 1024)         0           ['global_average_pooling1d_25[0][
                                                                 0]',
                                                                  'global_max_pooling1d_25[0][0]']

 dropout_125 (Dropout)          (None, 1024)         0           ['concatenate_25[0][0]']

 dense_125 (Dense)              (None, 512)          524800      ['dropout_125[0][0]']

 dropout_126 (Dropout)          (None, 512)          0           ['dense_125[0][0]']

 dense_126 (Dense)              (None, 512)          262656      ['dropout_126[0][0]']

 dropout_127 (Dropout)          (None, 512)          0           ['dense_126[0][0]']

 dense_127 (Dense)              (None, 256)          131328      ['dropout_127[0][0]']

 dropout_128 (Dropout)          (None, 256)          0           ['dense_127[0][0]']

 dense_128 (Dense)              (None, 128)          32896       ['dropout_128[0][0]']

 dropout_129 (Dropout)          (None, 128)          0           ['dense_128[0][0]']

 dense_129 (Dense)              (None, 20)           2580        ['dropout_129[0][0]']

==================================================================================================
Total params: 3,685,396
Trainable params: 1,685,396
Non-trainable params: 2,000,000
__________________________________________________________________________________________________
```

4. Conclusion

- Summary:

Test Summary (Weighted metrics)				
Model	Accuracy	Precision(W)	Recall(W)	F1-score(W)
SVM	0.66	0.68	0.66	0.66
BiLSTM	0.67	0.68	0.67	0.67

Thank you!