Incorporating Label Dependencies in Multilabel Stance Detection

William Ferreira
Lucubo Ltd
william@lucubo.com

Andreas Vlachos
Dept. of Computer Science and Technology
University of Cambridge
Cambridge, UK
av308@cst.cam.ac.uk

A Supplementary Material

This supplementary document contains additional details and plots omitted from the main paper.

A.1 Models

We use FastText v0.2.0, available from https://fasttext.cc/ (Joulin et al., 2017) as one of the baseline classifiers, for all 3 datasets. FastText is freely available general supervised learning classifier from Facebook's AI lab.

For the MTL models we construct a feed-forward, neural-network model, with a single hidden layer, regularised using dropout (Srivastava et al., 2014) with a fixed dropout rate of 0.5. We take as input the sentence embeddings of the utterances of each dataset, generated using an untuned version of the pre-trained ELMO (Peters et al., 2018) word embedding model[1] with dimension 1024. The MTL model is written using Keras (Chollet, 2015) and tensorflow (Abadi et al., 2016).

Consistent with (Simaki et al., 2017), the Logistic Regression (LR) classifier for the BBC dataset, uses the top 10 statistically significant linguistic features from (Simaki et al., 2018), which we augment with the top five hundred uni- and bi-grams, ordered by term frequency across the dataset, with English stop words removed, extracted using the sklearn (Pedregosa et al., 2011) CountVectorizer. The LR model is implemented in scikit-learn (Pedregosa et al., 2011).

A.2 Training

Accuracy, precision, recall and F1 (Sorower, 2010) are the standard metrics for binary classification tasks. The metrics can be extended to a multiclass setting by averaging over individual class scores. However, in a multilabel setting, the relationship between the true and predicted outputs is more complicated. Accuracy, also called *Exact Match Ratio* (Sorower, 2010) in a multilabel setting, can be a harsh metric, requiring, as it does, an exact match between the predicted and target labels. Consider the multilabel binary value $\mathbf{y} = [1, 0, 0, 1, 1]$, and prediction $\hat{\mathbf{y}} = [1, 1, 0, 0, 1]$. Since $\mathbf{y} \neq \hat{\mathbf{y}}$, the accuracy score for $\mathbf{y}$ is 0, however they partially match, since they are equal in the 2 out of 4 positions for which at least one is defined. Consequently, we use the *Jaccard Similarity Score* (JSS) (Jaccard, 1902), (Levandowsky and Winter, 1971), (Pedregosa et al., 2011) as our metric; the JSS for $\mathbf{y}$ and $\hat{\mathbf{y}}$ above is 0.5.

Hyper-parameter selection is done using 5-fold cross-validation using the relevant score metric. For the LR model the hyper-parameters are C: the inverse of regularisation strength, and the choice of regularisation penalty: L_1 or L_2. For the FastText models, the hyper-parameters are the word embedding dimension (300, 512, 1024), and the number of word ngrams to use (1, 2); the remaining FastText parameters are left at their default values. For the multi-task models, the hyper-parameters are the coefficients α coefficients controlling the contribution of the cross-label dependency loss to overall loss. We train the MTL models with stochastic gradient descent (SGD) using the Adam (Kingma and Ba, 2015) optimizer, and a batch size of 32, and 50 epochs of training.

A.3 Experiments

Figures 1, 2, 3, 4, 5, and 6 show the learning curves for the MFTC discourse domains omitted from the main paper. The learning curves show that for discourse domains ALM, BLM, Davidson, Election and MeToo, MTL-LP is superior to MTL-XLD at all training set sizes, however MTL-XLD is superior to MTL-LP for the Baltimore domain.

[1] https://tfhub.dev/google/elmo/2

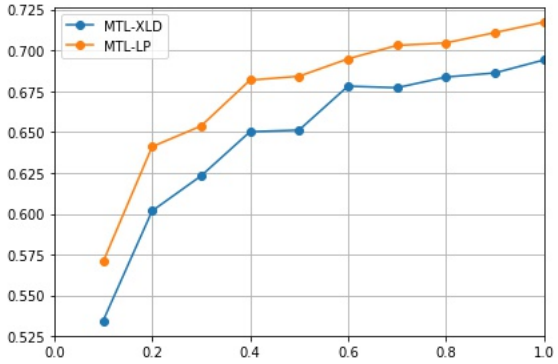

Figure 1: MFTC: ALM bootstrapped learning curve.

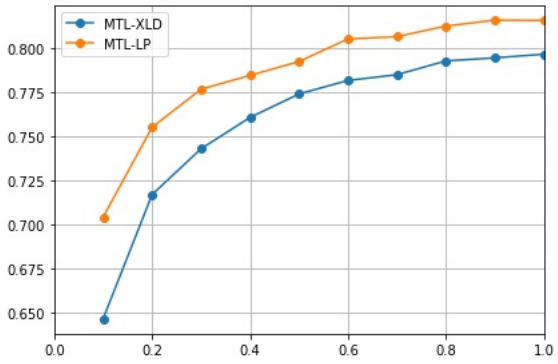

Figure 2: MFTC: BLM bootstrapped learning curve.

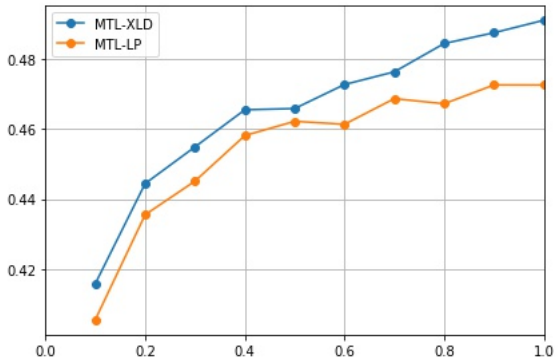

Figure 3: MFTC: Baltimore bootstrapped learning curve.

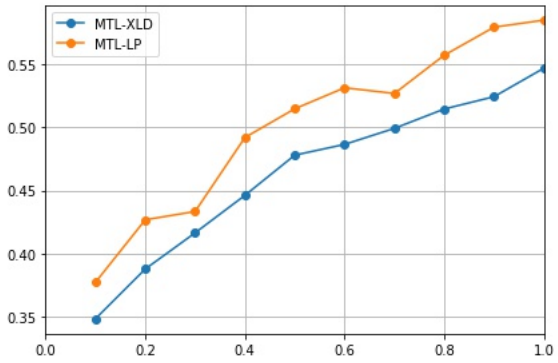

Figure 4: MFTC: Davidson bootstrapped learning curve.

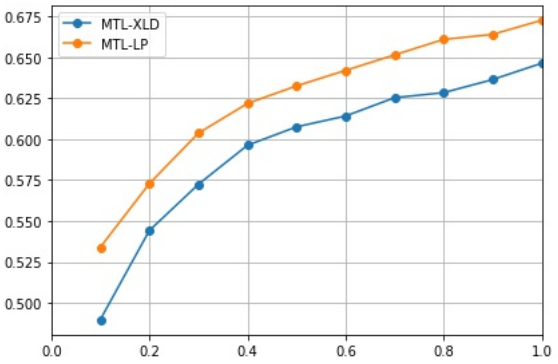

Figure 5: MFTC: Election bootstrapped learning curve.

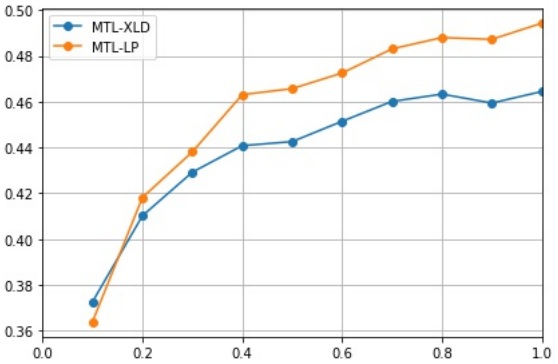

Figure 6: MFTC: MeToo bootstrapped learning curve.

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.

Franois Chollet. 2015. keras. `https://github.com/fchollet/keras`.

P Jaccard. 1902. Lois de distribution florale dans la zone alpine. *Bull Soc Vaudoise Sci Nat*, 38:69–130.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Michael Levandowsky and David Winter. 1971. Distance between sets. *Nature*, 234.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Vasiliki Simaki, Carita Paradis, and Andreas Kerren. 2017. Stance classification in texts from blogs on the 2016 british referendum. In *SPECOM*.

Vasiliki Simaki, Carita Paradis, and Andreas Kerren. 2018. Evaluating stance-annotated sentences from the brexit blog corpus: A quantitative linguistic analysis. *ICAME Journal*, 42:133–165.

Mohammad S Sorower. 2010. A literature survey on algorithms for multi-label learning. Technical report, Oregon State University, Corvallis.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.