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

Abstract

Stance detection in social media is a well-studied task in a variety of domains. Nevertheless, previous work has mostly focused on multiclass versions of the problem, where the labels are mutually exclusive, and typically positive, negative or neutral. In this paper, we address versions of the task in which an utterance can have multiple labels, thus corresponding to multilabel classification. We propose a method that explicitly incorporates label dependencies in the training objective and compare it against a variety of baselines, as well as a reduction of multilabel to multiclass learning. In experiments with three datasets, we find that our proposed method improves upon all baselines on two out of three datasets. We also show that the reduction of multilabel to multiclass classification can be very competitive, especially in cases where the output consists of a small number of labels and one can enumerate over all label combinations.

1 Introduction

Stance detection is an established task in the computational linguistics community, and is typically concerned with whether an utterance (e.g. a tweet) expresses an attitude (often *positive*, *negative* or *neutral*) against a target such as an entity e.g. a politician (Hasan and Ng, 2013; Mohammad et al., 2016), or another utterance, e.g. a previous tweet in a thread (Zubiaga et al., 2016). Thus stance detection is an important task for analyzing discourse in online forums and social media platforms and is a component in assessing the veracity of claims (Kochkina et al., 2018).

When the stances are mutually exclusive as in the aforementioned cases, multiclass classification is an appropriate formulation for the task. Often, however, a text may express multiple stances simultaneously. Such cases need to be formulated as multilabel classification (Sorower, 2010), where an instance can receive multiple, non-mutually exclusive labels. The most commonly used approaches to multiclass classification treat the task by learning models for each label. However, such approaches do not model dependencies between the labels explicitly, i.e. that the presence of one label results in another becoming more or less likely.

Brexit Blog Corpus (Simaki et al., 2018)
Utterance: rivalry between the us and china is inevitable but it needs to be kept within bounds that would preclude the use of military force.
Stances: certainty, contrariety, necessity, prediction
US Election Twitter Corpus (Sobhani et al., 2019)
Utterance: voters mean more than super delegates @sensanders corrupt -> #hillaryclinton spends millions on msm to discourage #americans voting #sanders
Stances: Clinton: AGAINST, Sanders: FAVOR
Moral Foundations Twitter (Dehghani et al., 2019)
Utterance: blatant racism in #colorado, #blacklivesmatter http://fb.me/1ibyxmswm
Stances: cheating, harm

Figure 1: Examples from each of the datasets.

In this paper, we investigate multilabel stance detection in the context of three datasets: the Brexit Blog Corpus (BBC) (Simaki et al., 2018), the US Election Tweets Corpus (ETC) (Sobhani et al., 2019), and the Moral Foundations Twitter Corpus (MFTC) (Dehghani et al., 2019). Figure 1 shows examples from each dataset where the utterances have been annotated with multiple stances. In BBC and MFTC, each utterance is annotated with a variable number of stances, encoded as binary presence/absence of each possible stance. In ETC, each utterance has a three-way stance *FAVOR* (positive), *AGAINST* (negative), or *NONE* (neutral) for each of the candidates.

We show that it is possible to improve over baseline results that employ binary relevance and

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6350–6354, Hong Kong, China, November 3–7, 2019.

multitask learning, by incorporating label dependencies explicitly with the *cross-label dependency loss* (Yeh et al., 2017), originally introduced by Zhang and Zhou (2006). We also show that a reduction of multilabel to multiclass classification by considering all label combinations, also known as label powerset, can be very competitive, especially when the output consists of a small number of labels and one can enumerate all combinations, and verify our results with statistical significance testing. Finally, we improve on the best reported results on the ETC dataset.

2 Multilabel classification

The most commonly used approach to multilabel classification encodes the labels so that a single multilabel classification task is reduced to many sub-tasks learned independently. E.g. for BBC and MFTC binary models are learnt for each of the labels that predict the presence or absence of each label, hence the name *Binary Relevance* (BR), which has also been used in image classification (Boutell et al., 2004). It is straightforward to extend BR to handle tasks such as ETC where each position in the output can have more options than presence/absence, by using multiclass classifiers instead of binary ones. While it is possible to learn the models for the subtasks jointly using multitask learning (Ruder, 2017), this does not capture label dependencies in the output directly; instead it encourages layers of the model before the output to be learned to benefit all tasks simultaneously.

An alternative encoding method, *Label Powerset* (LP), captures dependencies explicitly: each label combination appearing in the training data is encoded as a new, unique label, reducing the task once again to a multiclass classification. However, LP can introduce an exponentially large number of new labels, potentially with few training instances, thus exacerbating class imbalance. Moreover, only those label combinations seen in the data will be available during training; this can be a particular issue when there is a shortage of representative training data.

BR encoding methods ignore label dependencies, and the LP method relies on encoding each label combination appearing in the training set as an explicit new label, both methods reducing the task to binary/multiclass classification. In what follows, we adopt a middle-ground between BR and LP methods by incorporating a notion of dependence between the labels in the targets.

3 Cross-label dependency loss

To capture the dependencies among labels in the output, we follow Yeh et al. (2017) and employ the cross-label dependency (XLD) loss between vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$:

$$\text{XLD}(\mathbf{y}, \hat{\mathbf{y}}) := \frac{1}{|\mathbf{y}^0||\mathbf{y}^1|} \sum_{(p,q)\in\mathbf{y}^0\times\mathbf{y}^1} \exp(\hat{\mathbf{y}}_p - \hat{\mathbf{y}}_q)$$

where $\mathbf{y}$ denotes a vector of true (binary) labels of dimension n, $\hat{\mathbf{y}}$ a vector of predicted label probabilities, $\mathbf{y}^1$ are the indices of the 1-labelled elements of $\mathbf{y}$, $\mathbf{y}^0$ are the indices of the 0-labelled elements, and $\hat{\mathbf{y}}_p$ denotes the pth element of vector $\hat{\mathbf{y}}$. Minimising the cross-label dependency loss is equivalent to maximising the distance between 0- and 1-labelled targets, by penalising the model when it predicts label pairs that shouldn't co-exist for the instance. The intuition is similar to that of Bayesian Personalized Ranking in collaborative filtering (Rendle et al., 2009). We add XLD to cross-entropy loss to define an overall loss function $\text{L}(\mathbf{Y}, \hat{\mathbf{Y}})$:

$$\text{L}(\mathbf{Y}, \hat{\mathbf{Y}}) := \text{XEnt}(\mathbf{Y}, \hat{\mathbf{Y}}) + \alpha \sum_{i\in 1\ldots m} \text{XLD}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i)$$

where XEnt is cross-entropy loss, $\mathbf{Y}_i$ denotes a row vector of dimension n, and $\alpha \geq 0$ is a hyperparameter controlling the contribution of cross-label dependency loss to the overall loss.

Extending XLD to the ETC dataset is slightly complicated by the fact that the labels have three possible values, so we cannot represent a set of target labels as a binary vector. Firstly, we encode the labels using a *one-hot encoding* binary representation, so for example, $AGAINST = 100$, $NONE = 010$ and $FOR = 001$. We then apply XLD between the two encoded target labels, $\mathbf{y}$ and $\mathbf{z}$, of each tweet, and their predicted label probabilities $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ respectively, as follows:

$$\text{XLD}(\mathbf{y}, \mathbf{z}, \hat{\mathbf{y}}, \hat{\mathbf{z}}) = \sum_{(p,q)\in\mathbf{y}^0\times\mathbf{z}^1} \exp(\hat{\mathbf{y}}_p - \hat{\mathbf{z}}_q)$$

The above definition is not symmetric since it compares the 0-labelled positions of the first target label with the 1-labelled positions of the second target label. We re-introduce the symmetry by defining the overall loss function as:

$$\text{L}(\mathbf{Y}, \mathbf{Z}, \hat{\mathbf{Y}}, \hat{\mathbf{Z}}) := \text{XEnt}(\mathbf{Y}, \hat{\mathbf{Y}}) + \text{XEnt}(\mathbf{Z}, \hat{\mathbf{Z}})$$

$$+ \alpha_L \sum_{i \in 1...m} \text{XLD}(\mathbf{Y}_i, \mathbf{Z}_i, \hat{\mathbf{Y}}_i, \hat{\mathbf{Z}}_i)$$
$$+ \alpha_R \sum_{i \in 1...m} \text{XLD}(\mathbf{Z}_i, \mathbf{Y}_i, \hat{\mathbf{Z}}_i, \hat{\mathbf{Y}}_i)$$

where $\alpha_L \geq 0$ and $\alpha_R \geq 0$ are hyper-parameters controlling the contribution of the left and right XLD loss across the targets, respectively.

4 Experimental setup

In our experiments, we use the following multilabel stance detection datasets. The BBC dataset (Simaki et al., 2018) contains 1,239[1] utterances from social media blogs. Each utterance is assigned multiple stances by expert annotators from a set of ten stances. The ETC dataset (Sobhani et al., 2019) consists of 3 sets of tweets, each associated with a pair of election candidates in the US 2016 Election: Donald Trump-Hillary Clinton (DT-HC), Donald Trump-Ted Cruz (DT-TC), and Hillary Clinton-Bernie Sanders (HC-BS), containing 1,722, 1,317 and 1,366 tweets respectively. The MFTC dataset consists of 35,108 tweets curated from six[2] distinct discourse domains, e.g. natural disasters, politics, etc. Each tweet is annotated with up to 10 labels of moral sentiment.

Hyper-parameter selection is done using 5-fold cross-validation (CV) on the training set of each dataset. For the BBC dataset, we split the data 80% into a training set, and 20% holdout test set. For the ETC dataset, we combine the training and validation sets already provided to perform CV, and report on the original test set. For the MFTC data set, we split the data 80% into a training set, and 20% holdout test set.

In BBC and MFTC we use the *Jaccard Similarity Score* (JSS) (Jaccard, 1902) as our scoring metric:

$$J(X, Y) = \frac{\mathbf{y^1} \cap \hat{\mathbf{y}}^\mathbf{1}}{\mathbf{y^1} \cup \hat{\mathbf{y}}^\mathbf{1}} \quad (1)$$

JSS gives credit for partial matches, but does not reward predicting the absence of labels, which is desirable as most labels for each instance are absent (e.g. 90% of the instances in BBC and MFTC have at most two labels). It is less harsh than accuracy (*Exact Match Ratio*) (Sorower, 2010), which requires the entire label combination to be predicted correctly. For the ETC dataset, where each tweet is tagged with exactly two stances (i.e. no absent labels), following Sobhani et al. (2019), we use the macro-averaged F1-score for FAVOR and AGAINST, as the scoring metric.

5 Results

In our experiments, we consider models that capture label dependencies explicitly as well as baselines that do not capture these. As our baselines, we consider binary relevance using FastText (FT) classifiers (Joulin et al., 2017) for each stance label in BBC/MFTC and politician in ETC, as well as a multi-task learning (MTL) approach (Ruder, 2017) where each of the classifiers becomes a task and they all operate on a shared hidden layer (hard parameter sharing). As models capturing dependencies, we considered three options: the combination of the cross label dependency loss with MTL (MTL-XLD), and the combinations of label powerset with FT and MTL (FT-LP and MTL-LP respectively). For the latter, each label combination becomes a task learned jointly with the rest. Further details on all models and parameter tuning are in the supplementary material[3].

In Table 1 we report the test set results for all models. The results for the MFTC dataset are averaged across the six discourse domains. Overall, MTL-LP is the best performing multilabel classification method across all the datasets. MTL-LP is also better than the best performing model Seq2Seq reported in Sobhani et al. (2019) for the ETC dataset. MTL-XLD improves on the baseline models for the BBC and MFTC datasets, but performs slightly worse than MTL on the ETC dataset. We note that our results for the BBC and MFTC datasets are not directly comparable with previous work on BBC (Simaki et al., 2017) and MFTC (Dehghani et al., 2019), since we consider the full set of labels, whereas previous work removed those that were sparser. Our reimplementation of the logistic regression model of Simaki et al. (2017), as an additional baseline, resulted in poor performance in the BBC dataset (20 in JSS) and we did not consider it further.

5.1 Bootstrap training experiments

Reporting results on a held out test set is standard practice, but we also examine the results of

[1]The original dataset contained 1,682 utterances, but we removed duplicates occurring in the training and test sets.

[2]Originally seven but we dropped one domain after consultation with the authors.

[3]Code to reproduce our experiments is available here https://github.com/willferreira/multilabel-stance-detection

	BBC	ETC	MFTC
FT-BR	39.72	52.24	51.19
MTL	48.57	53.32	53.97
FT-LP	36.20	53.57	55.11
MTL-LP	**55.60**	**55.37**	**62.98**
MTL-XLD	51.33	52.22	60.94
Sobhani (Seq2Seq)	NA	54.81	NA

Table 1: Overall results for each dataset and model.

the classifiers for the BBC and MFTC datasets via bootstrapping (Gorman and Bedrick, 2019). We bootstrap 30 sample training sets from each dataset, by random sampling without replacement 80% of the data, retaining the 20% as a test set. For each sample, we train the MTL, MTL-LP, and best MTL-XLD model on each bootstrapped training set, and report the score on the test set. Summary statistics are shown in Tables 2 and 3. Using Welch's t-test[4], we cannot reject the null hypothesis (p=0.094) that the mean scores for the MTL-LP and MTL-XLD models on BBC are the same, but we can reject the null hypothesis that either of them are the same as the MTL baseline. For the MFTC Baltimore discourse domain, we cannot reject the null hypothesis (p=0.84) that the mean scores for the MTL-LP and MTL-XLD are the same, but we can reject the null that either of them are the same as the MTL baseline. For the remaining domains, we can reject the null hypothesis that the MTL-LP and MTL-XLD means are the same for ALM (p=6.74e-15), BLM (p=2.3e-13), Davidson (p=1.8e-24), Election (p=1.6e-17) and MeToo (p=0.0021). We can also reject the null, for each domain, that either MLT-XLD or MLT-LP have the same means as the MTL baseline.

	MTL	MTL-XLD	MTL-LP
mean	45.27	51.58	52.51
std	2.01	1.92	2.28

Table 2: Summary statistics for bootstrap results on BBC dataset.

5.2 Learning curve experiments

For the BBC and MFTC datasets, we construct the learning curve for the MTL-LP model against the MTL-XLD model. We sample increasingly larger fractions of the training sets, train the models on these fractions, and record the scoring metric on the original holdout test set. The learning curve for BBC is shown in Figure 2, from which we can see that the MTL-XLD model is superior to MTL-LP until the training dataset size is approximately 70% of the original, after which MTL-LP scores higher than MTL-XLD. The remaining learning curves for MFTC are given in the supplementary material, and 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.

[4] https://en.wikipedia.org/wiki/Welch%27s_t-test

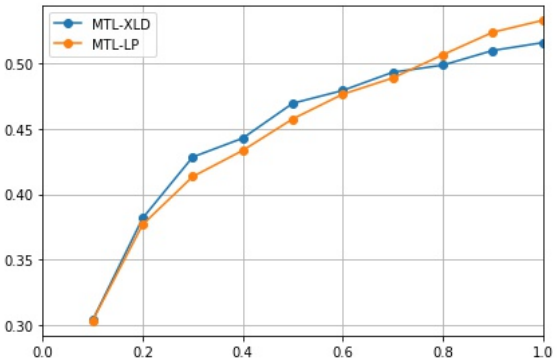

Figure 2: BBC: bootstrapped learning curve.

		ALM	BLM	B'more
MTL	mean	65.74	75.56	40.45
	std	1.71	1.11	1.82
MTL-LP	mean	**72.44**	**80.77**	48.45
	std	1.53	0.86	1.39
MTL-XLD	mean	69.82	78.03	**48.52**
	std	0.90	0.87	1.16
		D'son	Election	MeToo
MTL	mean	34.29	60.41	40.92
	std	3.41	1.78	1.55
MTL-LP	mean	**50.76**	**68.22**	**49.56**
	std	2.32	0.98	0.94
MTL-XLD	mean	48.76	65.43	48.75
	std	3.04	1.04	1.02

Table 3: Summary statistics for bootstrap results on MFTC dataset domains.

6 Conclusions

In this paper, we focused on multilabel stance detection and presented experiments on three

datasets. We demonstrated that taking label dependencies into account improves the performance of classification-based and multi-task baselines. In future work, we will explore how to integrate the textual descriptions of the labels in our our approach which has been shown to be beneficial in the case of large label sets (Mullenbach et al., 2018).

Acknowledgements

We would like to thank Daniel King for conducting initial experiments on the BBC corpus during his MSc thesis at the University of Sheffield.

References

Matthew Boutell, Jiebo Luo, Xipeng Shen, and Christopher Brown. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37:1757–1771.

Morteza Dehghani, Joseph Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Ying Lin, Aida M Davani, Brendan Kennedy, Mohammad Atari, Zahra Kamel, and et al. 2019. Moral foundations twitter corpus.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

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.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 452–461, Arlington, Virginia, United States. AUAI Press.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

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.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2019. Exploring deep neural networks for multi-target stance detection. *Computational Intelligence*, 35(1):82–97.

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

Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. 2017. Learning deep latent space for multi-label classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 2838–2844.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2438–2448, Osaka, Japan. The COLING 2016 Organizing Committee.