Does Multi-Task Learning Always Help? An Evaluation on Health Informatics Tasks

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, C Raina MacIntyre
CSIRO Data61, Sydney, Australia
Kirby Institute, University of New South Wales, Sydney, Australia
{firstname.lastname}@csiro.au , r.macintyre@unsw.edu.au

Abstract

Multi-Task Learning (MTL) has been an attractive approach to deal with limited labeled datasets or leverage related tasks, for a variety of NLP problems. We examine the benefit of MTL for three specific pairs of health informatics tasks that deal with: (a) overlapping symptoms for the same classification problem (personal health mention classification for influenza and for a set of symptoms); (b) overlapping medical concepts for related classification problems (vaccine usage and drug usage detection); and, (c) related classification problems (vaccination intent and vaccination relevance detection). We experiment with a simple neural architecture: a shared layer followed by task-specific dense layers. The novelty of this work is that it compares alternatives for shared layers for these pairs of tasks. While our observations agree with the promise of MTL as compared to single-task learning, for health informatics, we show that the benefit also comes with caveats in terms of the choice of shared layers and the relatedness between the participating tasks.

1 Introduction

Health informatics is the discipline concerned with the systematic processing of data, information and knowledge in medicine and healthcare (Hasman, 1998). Health informatics tasks tend to be specific in terms of parameters such as symptoms, regions of interest or the phenomenon to be detected. As a result, datasets for different health informatics tasks have been reported. However, it remains to be seen if these datasets or classification tasks help each other in terms of how similar the participating datasets or tasks are. In this paper, we examine the utility of Multi-Task Learning (MTL) for several pairs of health informatics tasks that are related in different ways.

MTL pertains to the class of learning algorithms that jointly train predictors for more than one task. In Natural Language Processing (NLP) research, MTL using deep learning has been used either to learn shared representations for related tasks, or to deal with limited labeled datasets (Xue et al., 2007; Zhang and Yeung, 2012; Søgaard and Goldberg, 2016; Ruder, 2017; Liu et al., 2017) for a variety of NLP problems such as sentiment analysis (Huang et al., 2013; Mishra et al., 2017). Most of this work that uses MTL presents architectures utilising multiple shared and task-specific layers. In contrast, we wish to see if the benefit comes from the simplistic notion of 'learning these classifiers together'. Therefore, we use a basic architecture for our MTL experiments consisting of a single shared layer and single task-specific layers, and experiment with different alternatives for the shared layer. This simplicity allows us to understand the benefit of MTL in comparison with Single-Task Learning (STL) for different configurations of shared layers, for task pairs that are related in different ways.

We experiment with datasets of English tweets for three pairs of boolean classification problems. The first pair deals with two datasets which were annotated for the same classification problem but differed in their scope in terms of illnesses that they cover. The second pair deals with different classification problems with some overlap in terms of the scope of medical concepts taken into account. The third pair deals with related classification problems: one problem influences the probability of output of the other.

Through our experiments with simple architectures for popular tasks in health informatics, we examine the question:
'*Does multi-task learning always help?*'

2 Related Work

MTL has been applied to a variety of text classification tasks (Søgaard and Goldberg, 2016; Xue et al., 2007; Ruder, 2017; Zhang and Yeung, 2012; Liu et al., 2017). The impact of task relatedness on MTL has been explored in case of statistical prediction models (Zhang and Yeung, 2012; Ben-David and Schuller, 2003). In the case of deep learning-based models, Bingel and Søgaard (2017) show how fundamental NLP tasks (such as MWE detection, POS tagging and so on) of different complexities perform when paired. (Mishra et al., 2017) use MTL for two related tasks in opinion mining: sentiment classificaton and sarcasm classification. (Wu and Huang, 2016) use MTL for personalisation of sentiment classification where global and local classifiers are jointly learned. A survey of MTL approaches using deep learning is by (Ruder, 2017).

In the context of health informatics, MTL has been applied in different kinds of tasks. Zou et al. (2018) predict influenza counts based on search counts for different geographical regions - however, they do not use a neural architecture. The task in itself is similar to Pair 1 in our experiments. Chowdhury et al. (2018) use MTL for pharmacovigilance, where each tweet is labeled with adverse drug reaction and indication labels. This is similar to the drug usage detection task in our experiments. For this, they use bi-directional LSTM as the shared layer, in addition to other task-specific layers before and after the shared layer. Benton et al. (2017) use MTL for prediction of mental health signals. Their architecture uses multi-layer perceptrons as shared layers. Bingel and Søgaard (2017) use bi-directional LSTM as the shared layer and compares different pairs of NLP tasks. In contrast, we experiment with three alternatives of shared sentence representations. The above are classification formulations for health informatics. MTL has also been used for other tasks such as biomedical entity extraction (Crichton et al., 2017), non-textual data based on medical tests to predict disease progression (Zhou et al., 2011) and so on.

We use datasets introduced in past work for our experiments. The sources of these datasets are described in the appropriate sections. In addition to the differences with past work as described above, to the best of our knowledge, the results of a MTL model have not been reported for the tasks and the task pairs that we consider. Our systematic analysis in terms of parameters of tasks and our experimentation with different shared layers sets us apart from past work.

3 Task Pairs Under Consideration

We consider three task pairs for our experimentation. These task pairs are related in different ways allowing an investigation into understanding configurations in terms of task relatedness in which MTL may be useful. The three configurations can be described as follows:

1. **Overlapping symptoms for the same classification problem**: The first pair corresponds to the same classification problem: personal health mention detection, *i.e.*, to predict if a given tweet reports an incidence of an illness, for overlapping concepts. For example, '*I have been sneezing since morning*' is a true instance, while '*Strong perfumes may cause sneezing*' is a false instance. Although the definition of the classification tasks is the same, we consider a pair of datasets that cover overlapping symptoms. The first dataset is labeled for personal health mentions of **influenza**, while the second dataset is labeled for personal health mentions of **multiple symptoms**, namely, cough, cold, fever and diarrhoea. Thus, this pair represents a configuration where the overarching classification task is the same but the set of multiple symptoms overlaps with the symptoms of influenza[1]. We refer to this as *Pair 1*.

2. **Overlapping medical concepts for different classification problems**: As *Pair 2*, we consider a pair of classification problems involving overlapping medical concepts. The tasks are: (1) **Vaccination behaviour detection**: To classify whether or not a person has received or intends to receive a vaccine; and, (2) **Drug usage detection**: To classify whether or not a person has received or intends to receive a medicinal drug. The relationship between these tasks arises because of the relationship between the medical concepts. For example, '*I got a flu shot yesterday*' is an instance of vaccine usage while '*I*

[1] https://www.cdc.gov/flu/consumer/symptoms.htm; Last accessed on 3rd September, 2019.

took a pain-killer yesterday' is an instance of drug usage. Since a 'vaccine' is a specific type of a general medical entity 'drug'[2], we expect that the classification tasks may be semantically different but deal with overlapping medical concepts.

3. **Related classification problems**: Finally, *Pair 3* corresponds to the configuration of related classification problems. The two classification problems that we consider are: (1) **Vaccine relevance detection**: To classify whether or not a tweet is relevant to vaccination; and, (2) **Vaccine intent detection**: To classify whether or not a tweet expresses intent to receive a vaccine. The classification tasks in pair 3 bear a notion of implication between them, because a tweet relevant to vaccines can alone express intent to receive a vaccine. For example, '*I don't think I will get a flu shot this year*' is relevant to vaccines but does not express vaccine intent.

For tasks in Pairs 1 and 2, we use datasets which contain labels for either of the tasks. Since the tasks in Pair 3 are related classification problems, each instance contains labels for both the tasks. The datasets were provided by three separate papers and may not contain purposeful overlaps.

4 MTL Architecture

We experiment with basic MTL architectures so as to understand the contribution of MTL to a fundamental architecture. The basic outline of our MTL architecture is shown in Figure 1. The input text is converted to a vector of embeddings using an embedding layer. This then goes to the shared sentence representation (hereafter referred to as the 'shared layer' for the sake of brevity), followed by the dropout layer. The dropout layer serves as an input to two dense layers, one for each classification task. The dotted rectangle in the architecture represents the shared layer. This layer is expected to capture the shared representation across the different classification tasks. In order to compare the role of different shared layers, we experiment with three configurations of neural layers as alternatives for the shared layer: BiLSTM, Convolutional, and BiLSTM followed by Convolutional. These are represented as B, C and $B + C$ in the rest of the paper.

In the case of Pairs 1 & 2, each instance carries values for exactly one of the two tasks because they were derived from two sources. Therefore, we consider it to be a case of missing labels. For each instance, we add a mask value of minus one (-1) for the label which is not present. We use a customised loss function which skips instances that bear the mask value. This means that instances that do not carry a label for a classification task are not incorporated when calculating the loss. For Pair 3, both labels are available for each instance. In this case, both labels are incorporated in computing the training loss.

5 Experimental Setup

All our tasks involve boolean text classification. We refer to the labels as 'true' and 'false' in the rest of the paper, although the semantics of these labels depend on the classification problem. We use the following datasets for our experiments:

- **Pair 1**:
  - As the '*influenza*' dataset, We use the dataset by Lamb et al. (2013) for influenza. The dataset contains 2,661 tweets (of which 1,304 are labeled as true). The original paper reports an n-gram baseline of 67%.
  - As the '*multiple symptoms*' dataset, we use a dataset of 9,006 tweets (of which 2,306 are labeled as true) by (Robinson et al., 2015). The tweets consist of illnesses, such as cough, cold, fever, and diarrhoea. No cross-validation results on this dataset have been reported in the original paper.
- **Pair 2**:
  - For vaccination usage detection, we use the dataset provided as a shared task as reported in Weissenbacher et al. (2018). The dataset consists of 5,751 tweets (of which 1,692 are true). The winning team by Joshi et al. (2018) reported a F-score of 80.87% for 10-fold cross-validation.
  - For drug usage detection, we use 13,409 tweets (of which 3,167 are true) provided by Jiang et al. (2016). No cross-

[2] https://www.cdc.gov/vaccines/vac-gen/imz-basics.htm; Last accessed on 3rd September, 2019.

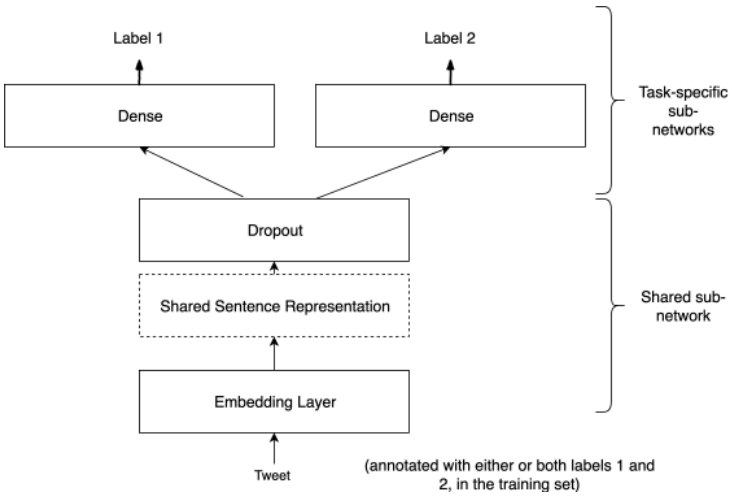

Figure 1: Our MTL architecture.

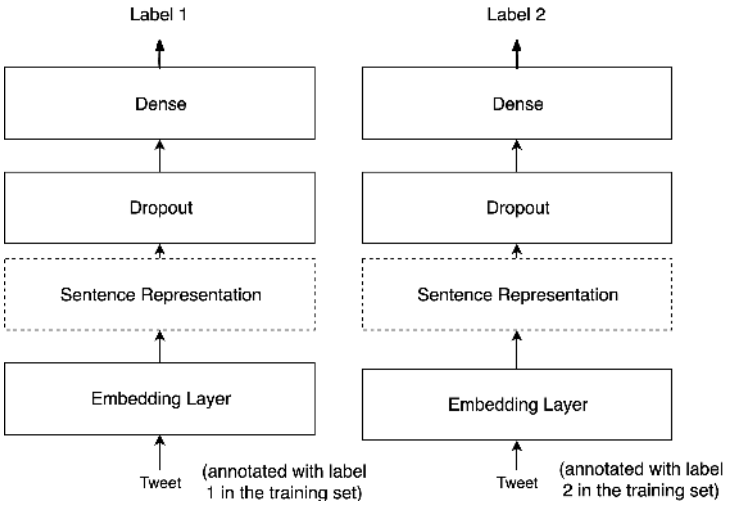

Figure 2: STL architecture corresponding to our MTL architecture.

validation results on this dataset have been reported in the original paper.

- **Pair 3**: We use a dataset of 10,688 tweets by Dredze et al. (2016). Out of these, 9,517 are labeled true for vaccine relevance while 3,097 are labeled true for vaccine intent. No experimental evaluation for these tasks has been reported in the paper or its derivative papers, to the best of our knowledge.

Since these datasets have been reported in past papers, we use Tweepy[3] to download the datasets of tweets using their identifiers. To implement the deep learning models, we use Keras (Chollet, 2015), with the Adam optimiser and binary cross-entropy as the loss function during training, with a dropout of 0.25 and number of units for intermediate layers as 25. We use word embeddings with 200 dimensions, pre-trained on a Twitter corpus using GLoVe (Pennington et al., 2014). These embeddings have been trained on 2 billion tweets with 27 billion tokens.

The general outline of our experimentation is a comparison of MTL with the equivalent single-task learning (STL) version. The corresponding STL architecture is shown in Figure 2. This architecture is identical to MTL, except that it separately learns the classifiers for the two tasks. The STL version uses one dense layer to obtain the classification output after the embedding layer and a layer to capture the semantic representation (the equivalent of the shared layer in MTL). For all our experiments, we report average accuracy and F-score values on ten-fold cross-validation.

[3] http://www.tweepy.org/; Last accessed on 3rd September, 2019.

	STL		MTL	
Shared Layer	Acc.	F-score	Acc.	F-score
	Influenza			
B	76.52	75.90	77.85	**76.41**
C	76.74	**76.75**	73.46	66.79
B+C	75.84	74.41	77.89	**76.86**
	Multiple symptoms			
B	78.49	48.48	75.34	**56.50**
C	74.91	**54.43**	79.58	44.29
B+C	78.39	45.34	79.28	**51.31**

Table 1: Accuracy and F-score (%) for Pair 1: Personal health mention detection for influenza and personal health mention detection for multiple symptoms.

	STL		MTL	
Shared Layer	Acc.	F-score	Acc.	F-score
	Vaccine Usage Detection			
B	85.46	74.85	85.90	**76.82**
C	85.53	**75.50**	85.59	75.47
B+C	84.28	73.49	85.62	**75.59**
	Drug Usage Detection			
B	78.78	53.74	79.27	**56.47**
C	77.20	**55.59**	80.74	54.59
B+C	78.09	52.79	80.71	**56.83**

Table 2: Accuracy and F-score (%) for Pair 2: Vaccination usage detection and drug usage detection.

	STL		MTL	
Shared Layer	Acc.	F-score	Acc.	F-score
	Vaccine Relevance Detection			
B	97.71	**98.80**	97.40	98.64
C	97.60	**98.75**	97.27	98.58
B+C	97.56	98.73	97.86	**98.88**
	Vaccine Intent Detection			
B	75.72	75.29	86.55	**78.93**
C	86.03	**76.93**	83.88	75.00
B+C	85.62	75.59	85.82	**77.15**

Table 3: Accuracy and F-score (%) for Pair 3: Vaccine relevance detection and vaccine intent detection.

6 Results

The effectiveness of Pair 1 for the three shared layers BiLSTM (B), Convolutional (C), and BiLSTM plus Convolutional (B+C) is shown in Table 1. These values are higher than the reported baseline for the influenza detection task. For both tasks, B and B+C result in an improvement when MTL is used. The highest improvement is 6% in case of influenza for B+C. However, there is a degradation when the shared layer is C. The improvement in case of B and B+C for 'Multiple symptoms' is statistically significant ($p < 0.05$, paired t-test). The improvement in the case of influenza, however, is not statistically significant.

The corresponding effectiveness of Pair 2 is shown in Table 2 for the pair: vaccine usage detection and drug usage detection. The best F-score for vaccine usage detection is 76.82%, when a BiLSTM layer is used as a shared representation in the MTL architecture. The best F-score for drug usage detection is 56.83%, when a combination of BiLSTM and convolutional layers is used in a corresponding setting. We observe that, for vaccine usage detection, there is an improvement of 2-3% in case of either B or B+C. The improvement is not statistically significant. Similar trends are observed for Pair 3, as shown in Table 3. We observe that the F-scores are also high (around 97-98%) for vaccine relevance detection, purely due to the skew in the dataset. However, we observe that, in this case, the improvement in F-score when MTL is used is observed only in the case of B+C. Since vaccine intent *implies* vaccine relevance, our results show that MTL may not be beneficial for related classification tasks where one task implies another.

It is possible that the benefit of MTL depends on the size of the training set from the conjugate task (*i.e.*, tweets labeled for drug usage detection in order to improve the effectiveness of vaccine usage detection). Therefore, the impact of the size of training dataset on accuracy of four tasks of pairs 1 and 2, for the best performing architecture is shown in Figure 3. In general, the performance improves with an increase in training set size. The improvement is higher for Pair 2 than Pair 1. The datasets in Pair 2 were created using separate sets of keywords (drug names versus vaccine names, in specific), while the ones in Pair 1 were created using overlapping sets of keywords. Thus, the extent of relatedness or similarity governs the performance gain due to MTL. It may be noted that this comparison is not relevant for Pair 3 since every

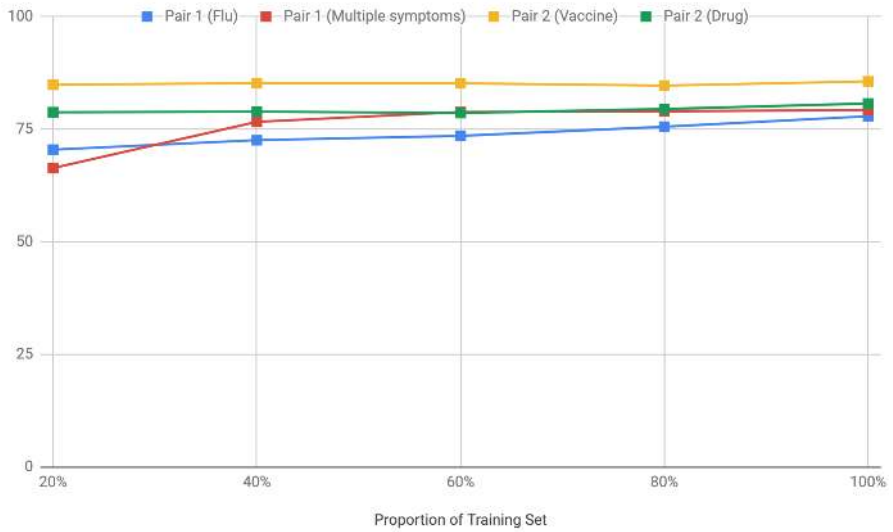

Figure 3: Change in accuracy values with an increase in the proportion of training set from the additional task in the pair for pairs 1 and 2.

instance in the dataset contains both the labels.

7 Error Analysis

We analyse the errors in two parts. In the first part, we compare errors made by the architectures that use STL and MTL. This helps to understand situations in which MTL does better than STL. In the second part, we evaluate errors made by MTL. These can serve as pointers for future work.

We manually analyse 50 randomly selected erroneous instances each from STL and MTL for all pairs of tasks. The benefit of MTL over STL was observed in the following cases for Pairs 1 and 2:

- **Pair 1**: For personal health mention detection, false positives were observed in the form of tweets that express the fear of flu (for example, the tweet '*i feel like im getting sick!=[UGH piggy flu stay away!*' was mis-classified) in the case of STL but not in MTL. Errors due to figurative language (for example, '*because theres times when i want to just check my facebook feed and not feel sick to my stomach*') occurred with STL (6 out of 50) more often than with MTL (2 out of 50).

- **Pair 2**: Errors in tweets where the speaker was reluctant to take a drug (for example, the tweet '*do i take my migraine medicine and pray for no interactions or do i take a muscle relaxant or tramadol and hope for the best*' was mis-classified) were reduced when MTL (1 out of 50) was used instead of STL (8 out of 50).

We observe no specific patterns of errors for Pair 3 when we compare the mis-classified instances for STL and MTL.

In contrast, an analysis of errors obtained from MTL showed the following patterns:

- Errors in **Pair 1** include long tweets which contain a rant along with a personal health mention (11 out of 50). For example, '*8 hrs sleep still feel like shit laying in pitch black listening to my belly make some weird arsed noises think im gunna hurl again*'.

- Errors in **Pair 2** include (a) Apprehensions/fears expressed before a flu shot/Intent to receive a flu shot (16 out of 50, in case of vaccination usage detection); (b) Mentions of a drug for dramatic effect (14 out of 50, in case of drug usage detection). For example, '*i dont usually remember drunk dreams. unless combined w melatonin*'.

These show that MTL may be unable to guard against topic drifts observed due to rants, apprehensions or dramatisation.

8 Conclusions & Future Work

We evaluate multi-task learning (MTL) for three pairs of similar health informatics tasks dealing with: (1) Overlapping symptoms (detection

of influenza and multiple symptoms); (2) General/specific medical concepts (detection of the usage of drugs and vaccines); and, (3) Related classification problems (vaccine relevance detection and vaccine intent detection). We compare STL with MTL where the pair of tasks are jointly learned for three kinds of shared sentence representations. In general, for shared layers based on BiLSTM and BiLSTM + Convolutional, MTL helps the three pairs. However, this improvement is not observed when the Convolutional layer is used as a shared representation. The improvement, wherever applicable, is around 2-4% for all the pairs. While MTL has been considered almost a 'silver bullet' in situations where related classification problems or datasets are available, our results highlight the caveats therein. We observe that the benefit of MTL depends on the type of shared layer and the relationship between the tasks under consideration.

Our results show that MTL can help to leverage different datasets annotated for related health informatics tasks. This is potentially useful since specialised tasks are common in health informatics and large datasets may or may not be available. It remains to be verified if the benefit can be generalised for other tasks. Similarly, while we present the relatedness between the participating tasks in a qualitative manner, their similarity could be empirically determined as a future work. A correlation between the similarity of classification tasks and the expected benefit of MTL is a possible future work.

References

Shai Ben-David and Reba Schuller. 2003. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–162, Valencia, Spain.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 164–169, Valencia, Spain.

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

Shaika Chowdhury, Chenwei Zhang, and Philip S Yu. 2018. Multi-task pharmacovigilance mining from social media posts. In *Proceedings of the World Wide Web Conference*, pages 117–126, Lyon, France. International World Wide Web Conferences Steering Committee.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368.

Mark Dredze, David Broniatowski, Michael Smith, and Karen M Hilyard. 2016. Understanding vaccine refusal: why we need social media now. *American journal of preventive medicine*, 50(4):550–552.

Arie Hasman. 1998. Education and health informatics. *International journal of medical informatics*, 52(1-3):209–216.

Shu Huang, Wei Peng, Jingxuan Li, and Dongwon Lee. 2013. Sentiment and topic analysis on social media: A multi-task multi-label classification approach. In *Proceedings of the Annual ACM web science conference*, pages 172–181, Paris, France.

Keyuan Jiang, Ricardo Calix, and Matrika Gupta. 2016. Construction of a personal experience tweet corpus for health surveillance. In *Proceedings of the ACL Workshop on biomedical natural language processing*, pages 128–135, Berlin, Germany.

Aditya Joshi, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2018. Shot or not: Comparison of nlp approaches for vaccination behaviour detection. In *Proceedings of the EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 43–47.

Alex Lamb, Michael Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, Atlanta, Georgia.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1–10, Vancouver, Canada.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 377–387, Vancouver, Canada.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1532–1543, Doha, Qatar.

Bella Robinson, Ross Sparks, Robert Power, and Mark Cameron. 2015. Social media monitoring for health indicators. In *International Congress on Modelling and Simulation*, Gold Coast, Australia.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 231–235, Berlin, Germany.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In *Proceedings of the EMNLP Workshop on The Social Media Mining for Health Applications Workshop and Shared Task*, pages 13–16, Brussels, Belgium.

Fangzhao Wu and Yongfeng Huang. 2016. Personalized microblog sentiment classification via multi-task learning. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*.

Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. 2007. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8(Jan):35–63.

Yu Zhang and Dit-Yan Yeung. 2012. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*.

Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. 2011. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822. ACM.

Bin Zou, Vasileios Lampos, and Ingemar Cox. 2018. Multi-task learning improves disease models from web search. In *Proceedings of the World Wide Web Conference*, pages 87–96, Lyon, France. International World Wide Web Conferences Steering Committee.