Multi-task Learning Framework for Mining Crowd Intelligence towards Clinical Treatment

Shweta Yadav*, Asif Ekbal*, Sriparna Saha*, Pushpak Bhattacharyya*, and Amit Sheth†
*Indian Institute of Technology Patna, India
†Kno.e.sis Center, Wright State University, USA
*{shweta.pcs14,asif,sriparna,pb}@iitp.ac.in, †amit@knoesis.org

Abstract

In recent past, social media has emerged as an active platform in the context of healthcare and medicine. In this paper, we present a study where medical user's opinions on health-related issues are analyzed to capture the medical sentiment at a blog level. The medical sentiments can be studied in various facets such as *medical condition*, *treatment*, and *medication* that characterize the overall health status of the user. Considering these facets, we treat analysis of this information as a multi-task classification problem. In this paper, we adopt a novel adversarial learning approach[1] for our multi-task learning framework to learn the sentiment's strengths expressed in a medical blog. Our evaluation shows promising results for our target tasks.

1 Introduction

"Can someone help me please????". These types of queries have swamped the web with the phenomenal rise in social media contents almost in every domain including health care. Generally, the users posts seeking for health-related information, sharing medical experiences and opinions of other users (i.e., patients, health professional or doctors). With the enormous amount of posts increasing day after day, it is difficult for the health professionals to read and answer every post. It would be helpful to have a sentiment analyzer that could study the user's sentiment associated with the post related to his/her health-status. In this paper, we make an attempt to capture *medical sentiment* (MS) by analyzing the subjectivity expressions describing a patient's medical conditions at the blog level. MS can be studied as an event that characterizes the patient's medical condition, in which the patient expresses stance towards clinical and social situations. The notion of sentiment in medical context unlike traditional sentiment analysis (SA) is more granular which can be studied after considering various aspects (Denecke and Deng, 2015) that can directly impact the users' health conditions, such as:

(1) Changes in the medical condition (e.g., *Sentiment can be observed as a change in a patient's medical condition which can improve or worsen over a time.*)

(2) Severity of the medical condition that impacts patient life (e.g., *severe headache impacts the patient's life more than a mild headache.*)

(3) Outcome of a treatment (e.g., *there may be positive or negative impacts in a patient's treatment.*)

In the current study, the problem of medical sentiment identification is addressed by exploiting two important associated aspects as shown in Figure-1 and leveraging their synergies in a deep multi-task learning framework. In recent years, neural network models have gained their popularity for solving problems in several domains (Misra et al., 2016; Luong et al., 2015), as they facilitate an efficient way of amalgamating information from several tasks. This method of multi-task learning provides advantages in (1) minimizing the number of parameters and (2) reducing the risk of over-fitting. The aim of multi-task learning (MTL) is to efficiently enhance the system performance by integrating the other similar tasks. The primal factor of MTL is the sharing scheme in latent feature space. Most of the existing methods on multi-task classification attempt to divide the features of different tasks based on task-specific and task-invariant feature space, considering only parameters of some components that could be shared. The major drawback of this mechanism is that the common feature space often incorporates some redundant task-variant

[1]The reader is encouraged to contact the authors regarding the availability of data and source code

Proceedings of NAACL-HLT 2018, pages 271–277
New Orleans, Louisiana, June 1 - 6, 2018.

Task1: Medical Condition			Task2: Medication		
Medical Blog-text	**Label**	**Description**	**Medical Blog-text**	**Label**	**Description**
*"I felt an **incredible surge of unsteadiness**"*	Exist	User shares the **symptoms (negative sentiment)** of any medical problem.	"Hi been on Sertaline now for abut 4 weeks. My mood has **definitely improved**"	Effective	User shares the **positive sentiment** in the form of **usefulness of the treatment.**
*"Previously I have taken flixonase which has given no long term relief 10 days ago I went back to the doctor and was given Betnesol. This has **immediately relieved me all symptoms**"*	Recover	User shares the **recovering status (positive sentiment)** from the previous medical problems.	"Had anxiety for few months on 2mg diazapam. Nothing seems to help been in bed for two days **cant sleep waking at 3 4 5 am**"	Ineffective	The **no effect** of the **treatment** is reported in the user narration.
*"I recently started lexapro 3 days, Im absolutely lost I feel **weak and shaky everyday** and **cant eat right I dont sleep normal. Ill die young** and thecause will be cardiac arrest"*	Deteriorate	User describes its medical condition to be **worsen (negative sentiment)** over the span of medical treatment.	"Day 6 that I have taken my citalopram. Anxiety is down, but now Im starting to **feel more and more off.**. Random **high chest pains** Plus feeling a **bit foggy and spacey.**."	Serious Adverse Effect	User shares the **negative opinion** towards the treatment mainly in the form of adverse drug effect.

Figure 1: Exemplar description of medical blog-text from two different medical aspects (medical condition and medication). The texts in bold indicates the sentiment word.

features, while certain common features could also lie in the task specific feature space, leading to feature redundancy.

Adversarial learning (Goodfellow et al., 2014) is the process of learning a model to correctly classify both unmodified data and adversarial data through the regularization method. It can be used to combat this issue by ensuring the mutual representation between the task that could inherently disjoint task-specific and task-invariant feature space. This helps in eliminating redundant features from the feature space.

Motivated by the success of adversarial learning in several classification tasks (Miyato et al., 2016; Ge et al., 2017), we adopt the adversarial multi-learning framework to capture the MS in various medical aspects.

Contribution: **(i)** a description of the medical-sentiment classification task by mining medical blogs using users sentiments towards medical condition and medication, and **(ii)** a method for analysis of medical sentiments over various aspects by exploiting the multi-task adversarial training framework which enables multiple aspects of MS tasks to be jointly trained.

2 Related Works

In the recent past, there has been a significant growth in the studies to analyze the sentiment of users in a healthcare/medical domain. The study conducted by Denecke and Deng (2015) provides the quantitative assessment of sentiment across the clinical narrative and social media sources. Towards this, they created a domain-specific corpus from MIMIC II database containing clinical documents (nurse letters, radiology reports, and discharge summaries). They also studied users self reported drug reviews on blogs (WebMD, DrugRating) to asses the possible medical sentiments. Majority of the current research in medical sentiment analysis are focused on understanding the mental health disorder, mainly depression. Several shared tasks (Losada et al., 2017; Hollingshead et al., 2017) have also been organized to study the patient health-related opinions on social media. The challenge defined in Milne et al. (2016) aims to automatically classify the user posts from an online mental health forum into four different categories (crisis/red/amber/green) according to how urgently the post needs the attention.

Shickel et al. (2016) introduced the notion of applying sentiment analysis to the mental health domain by defining new polarity classification scheme. They split the traditional 'neutral' class into both a dual polarity sentiment (both positive and negative) and a 'neither positive nor negative' sentiment class. Some of the other prominent works in the opinion mining in medical setting, includes studies by (Bobicev et al., 2012; Sokolova and Bobicev, 2011; Ali et al., 2013).

In the study conducted by (Pestian et al., 2012), authors analyzed the emotions and sentiment of suicide notes. The other study in medical sentiment analysis includes the work of Bobicev et al. (2014), where they analyzed sequences of sentiments (encouragement, gratitude, confusion, facts, and endorsement) in In Vitro Fertilization (IVF) medical forum.

In terms of methods, majority of the work utilizes machine learning technique (SVM, naive Bayes, logistic regression) by exploiting features such as

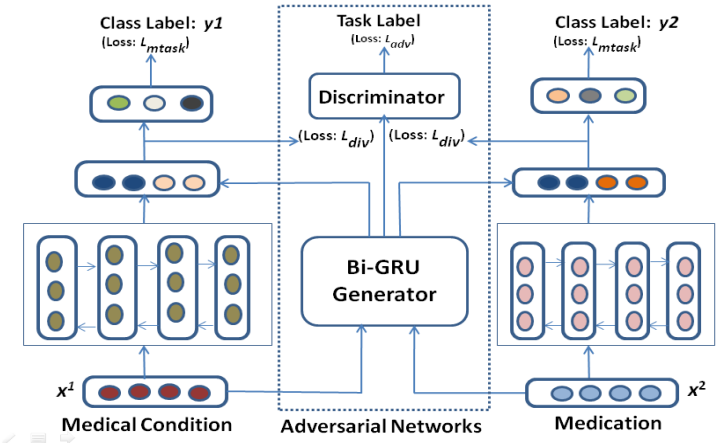

Figure 2: Architecture of proposed methodology

bigram, trigram, parts of speech. Also, there has been predominant use of general sentiment lexicon, however their analysis shows that it does not help in capturing the medical sentiment. More domain specific knowledge is also embedded using medical knowledge graph such as UMLS to identify the medical condition and treatment (Sokolova et al., 2013).

3 Overview of the proposed model

We formulate the MS analysis problem as a multi-task classification problem.

Problem Statement: Let us assume that a blog-text P consisting of k sentences i.e., $P = \{s_1, s_2 \dots s_k\}$ and the set of tasks, $T = \{t_1, t_2\}$ be given. Let the data set of task $t \in T$ be $\mathscr{D}_t = \{(x_t^n, y_t^n) : n = 1 \dots N_t\}$, where x_t^n denotes a blog-text P with the corresponding label y_t^n from a task t having N_t instances. The task is to predict $\bar{y}_t$ such that $\bar{y}_t = \text{argmax}_{y_t}\{p(y_t|x_t)\}$.

We clearly illustrate the two tasks related to MS identification in Figure-1.

In this section, we present an overview of the proposed model for multi-task medical sentiment classification. We use the bi-directional gated recurrent units (Bi-GRU) (Chung et al., 2014) to encode the blog-text as it is computationally cheaper than long short term memory (LSTM) (Hochreiter and Schmidhuber, 1997). The updates for Bi-GRU units can be computed by

$$h_t = BiGRU(h_{t-1}, x_t) \tag{1}$$

where, h_t and h_{t-1} are the hidden units at time t and $(t-1)$, respectively. x_t is the input at time t.

3.1 Classification of Medical Blog

Let us assume that a blog-text P having k sentences and word sequence $w = \{w_1, w_2, \dots w_l\}$ be given. The embedding layer is used to find out the vector representation $x_i \in \mathcal{R}^{d \times V}$ from a d dimensional pre-trained word embedding of vocabulary V. Each word $w_i \in w$ will be represented by its respective word embedding x_i . The hidden units h_l learned at the last time step (l) of sequence are considered as the encoding of the medical blog, P. The representations h_l generated from the Eq 1 are fed to a fully connected softmax layer to generate the probability distribution over the given classes.

$$\bar{y} = softmax(h_l^T W + z) \tag{2}$$

Here, W and z are weight matrix and bias vector, respectively. The term $\bar{y}$ denotes the predicted probability distribution.

Loss Function: Cross entropy is used to define the loss function. Given a training dataset $\mathscr{D} = \{(x^i, y^i) : i = 1 \dots N\}$, the network parameters are trained to minimize the cross entropy of the predicted probability distributions $(\bar{y})$ and true probability distributions (y) over the C number of classes.

$$\mathcal{L}(y, \bar{y}) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j log(\bar{y}_i^j) \tag{3}$$

The above loss function can be extended for our problem in the following ways:

$$\mathcal{L}_{mtask} = \lambda_1 \mathcal{L}(y_1, \bar{y}_1) + \lambda_2 \mathcal{L}(y_2, \bar{y}_2) \tag{4}$$

where λ_1 and λ_2 are the weight factors for the task 'Medical condition' and 'Medication', respectively.

Task 1: Medical Condition					
Total blog-post	Exist	Recover	Deteriorate	Avg # of sentences	Avg # of words
5188	2396	703	2089	10	192
Task 2: Medication					
Total blog-post	Effective	Ineffective	Serious Adverse Effect	Avg # of sentences	Avg # of words
2301	462	613	1,226	9	176

Table 1: Dataset statistics for Task 1 and Task 2

Features for multi-task learning: The multi-task learning is governed by sharing the latent features over different tasks. In the proposed neural network based model, the features are the hidden states of BiGRU at the end of sequence. Motivated by the shared-private feature sharing scheme in (Liu et al., 2017), for each task we define two feature spaces; task-specific and task-invariant. Mathematically, for a given blog-text P of task t, we can compute its task-specific features $h_l^t = BiGRU(h_{l-1}^t, x_l)$ and task-invariant features $f_l^t = BiGRU(f_{l-1}^t, x_l)$. Subsequently, the final features will be the concatenation of both features.

3.2 Adversarial Training

Although the feature sharing scheme separates the features into two features spaces, but there is no guarantee that contamination will not be made. Inspired by adversarial networks, we follow the generative-discriminative strategy to avoid the contamination in features space in which a BiGRU works as generator (G) to generate task-invariant features. A discriminator model (D) is used to map the task-invariant features of a blog-text into a probability distribution. It is mainly a multi-layer perceptron classifier which classifies a blog sentence into its respective tasks. The adversarial loss is used to train the model which produces task-invariant features such that a classifier cannot reliably predict the task based on these features. Similar to (Goodfellow et al., 2014; Liu et al., 2017), we use the following adversarial loss function

$$\mathcal{L}_{adv} = \min_{G} \Big(\max_{D} \big(\sum_{t=1}^{T} \sum_{i=1}^{N_t} d_i^t log[D(G(x^t))] \big) \Big) \quad (5)$$

where d_i^t is the gold label indicating the type of the current task. Based on the recent work (Bousmalis et al., 2016; Liu et al., 2017) on shared-private latent space analysis, we introduce another divergence loss function $\mathcal{L}_{div}$ to castigate the redundant features and encourage the task-invariant and task-specific feature extractors to encode different aspects of the inputs. The divergence loss function can be computed as $\mathcal{L}_{div} = \sum_{t=1}^{T} \|F^{t^T} H^t\|_F$, where F^t and H^t are two matrices, where rows are task-invariant and task-specific features of a blog-text from a task t. The $\|.\|_F$ denotes the Frobenius norm of the matrix. The final loss function $\mathcal{L} = \alpha_1 \mathcal{L}_{mtask} + \alpha_2 \mathcal{L}_{adv} + \alpha_3 \mathcal{L}_{div}$ is used as underlying loss function to train the network. Here α_1, α_2 and α_3 are the hyper-parameters of the networks.

4 Dataset and Experimental Setup

We generate a corpus[2] of $7,490$ blog-text collected on four popular groups, namely *Depression, Allergy, Asthma, and Anxiety*. Out of total blog-text, $5,188$ blogs concern about the *medical conditions* and $2,302$ are classified as *medication*. We provide the detailed dataset statistics for both the task is presented in Table-1. A team of three annotators [3] independently annotated the user posts with three classes on both the classification strategies. The Cohen's kappa approach (Cohen, 1960) was used to measure the inter-annotator agreement. We observe high agreement ratio of 0.79 (task 1) and 0.84 (task 2) for exact matching of the class w.r.t each blog post. We have performed 5-fold cross-validation experiment on both the datasets. The pre-trained embeddings (Mikolov et al., 2013) of dimension 300 were used in the experiments. The dimension of Bi-GRU hidden unit is set to 100 via grid search, on the basis of cross-validation performance. We choose the same value of 0.5 for both the weight factors λ_1 and λ_2 to impose equal importance on both the tasks. Training was performed using stochastic gradient descent over mini-batches of size 50 considering the Adadelta (Zeiler, 2012) update rule with an initial learning rate of 0.01. The min-max optimization is performed with the help of gradient reversal layer (Ganin et al., 2016). As a regularizer, we use dropout (Hinton et al., 2012) with a probability of 0.5. We train the network with 130 epochs. The optimal [4] hyper-parameter values are obtained via a grid search for α_1, α_2, and α_3 over the best cross-validation performance.

[2] Accepted at LREC-2018; entitled "Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System" and will be publicly available.

[3] undergraduate students with the medical knowledge

[4] $\alpha_1 = 0.88$, $\alpha_2 = 0.12$ and $\alpha_3 = 0.03$

Models	Task 1: Medical Condition			Task 2: Medication		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Baseline 1: MT-LSTM	63.40	61.38	62.37	88.23	77.38	82.45
Baseline 2: ST-LSTM	63.19	62.47	62.83	85.94	77.46	81.48
Proposed Approach	66.82	63.61	65.18	85.83	81.79	83.76

Table 2: Performance comparisons of our proposed approach with baselines.

4.1 Performance Evaluation

In order to show the effectiveness of our proposed method, we chose the neural network models popular in single task and multitask setting for our specified problem of text classification.
Baseline 1: Single Task-LSTM
Baseline 2: Multi Task-LSTM (Liu et al., 2016).
Table-2 reports the results of our proposed approach with baselines system. From the results, we observe that the performance on both the tasks significantly increase with the introduction of adversarial learning in multi-task framework. More specifically, compared to baseline 1, we observe the performance improvement of 2.81 and 1.31 F-score points on Task 1 & 2, respectively. In multi-task framework (Baseline 2), our system achieves the improvements of 2.35 and 2.28 F-score points on Task 1 & 2, respectively. We also analyze that mere introduction of multi-task framework sometimes may cause a drop in performance. This is because of the shared feature-space which includes both private and shared features leading to redundancy. Statistical significance test shows that the improvements over both the baselines are statistically significant as (p-*value* < 0.05).

4.2 Analysis

Our analysis on medical blog-text discovers that unlike traditional SA study on social media text, SA on medical text owes several unique challenges which have formed the major causes of the errors:
(1) Usually, the user present the health related information in a more elusive way which requires deeper analysis of metaphor and sarcasm. For example:
" Lol I'm just a big ball of anxiety fun.",
" My head is like air."
(2) MS is often presented implicitly which need to be inferred, for instance, from the medical concepts used in documents. Implicit MS (Exist) present in the blog are for example:
"It almost feels like im half awake and half asleep."
(3) The usage of abbreviated and short words have become ubiquitous in medical blog text. For e.g.,***"Cit" for the "Citopram".***
(4) The context scope of a sentiment changes extensively from a single phrase to multiple sentences. Moreover, adversative transitive words were widely used to link these phrases or sentences. The medical sentiment was bounded and implied by these inter and intra- sentence discourse relations. For example:
"The thoughts are of anything which is quite good. I have an anxiety disorder but I can't cope with it..."

5 Conclusion and Future Work

In this paper, we have introduced different aspects of sentiments in the context of medicine such as 'medication' and 'medical condition' instead of conventional polarity to judge user's health status. For this, we have utilized highly representative medical blog text to validate our study. We have proposed a robust sentiment-sensitive multi-task framework, settling on adversarial learning to capture the medical sentiment in the user's blog-post. We were able to obtain significant performance improvements over the state-of-the-art baseline system in all the cases. In future, we plan to address the implicit and sarcastic medical sentiments that account to the majority of the errors.

Acknowledgement

Asif Ekbal acknowledges Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia). Amit Sheth acknowledged partial support from NMH award R01MH105384 "Modeling Social Behavior for Healthcare Utilization in Depression. All findings and opinions are of authors and not sponsors.

References

Tanveer Ali, Marina Sokolova, David Schramm, and Diana Inkpen. 2013. Opinion learning from medical forums. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. pages 18–24.

Victoria Bobicev, Marina Sokolova, Yasser Jafer, and David Schramm. 2012. Learning sentiments from tweets with personal health information. In *Canadian Conference on Artificial Intelligence*. Springer, pages 37–48.

Victoria Bobicev, Marina Sokolova, and Michael Oakes. 2014. Recognition of sentiment sequences in online discussions. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*. pages 44–49.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems*. pages 343–351.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* .

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

Kerstin Denecke and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine* 64(1):17–27.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17(59):1–35.

ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. 2017. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418* .

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. pages 2672–2680.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* .

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Kristy Hollingshead, Molly E. Ireland, and Kate Loveys, editors. 2017. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Vancouver, BC. http://www.aclweb.org/anthology/W17-31.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101* .

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1–10. http://aclweb.org/anthology/P17-1001.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations .

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114* .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. pages 118–127.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 3994–4003.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725* .

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights* 5:BII–S9042.

Benjamin Shickel, Martin Heesacker, Sherry Benton, Ashkan Ebadi, Paul Nickerson, and Parisa Rashidi. 2016. Self-reflective sentiment analysis. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. pages 23–32.

Marina Sokolova and Victoria Bobicev. 2011. Sentiments and opinions in health-related web messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. pages 132–139.

Marina Sokolova, Stan Matwin, Yasser Jafer, and David Schramm. 2013. How joe and jane tweet about their health: Mining for personal health information on twitter. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. pages 626–632.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR* abs/1212.5701. http://arxiv.org/abs/1212.5701.