Transformer-based Multi-Task Learning for Adverse Effect Mention Analysis in Tweets

George-Andrei Dima[1,2], Dumitru-Clementin Cercel[1], Mihai Dascalu[1]
University Politehnica of Bucharest, Faculty of Automatic Control and Computers[1]
Military Technical Academy Ferdinand I[2]
andrei.dima@mta.ro, {dumitru.cercel, mihai.dascalu}@upb.ro

Abstract

While social media gains a broader traction, valuable insights and opinions on various topics representative for a wider audience can be automatically extracted using state-of-the-art Natural Language Processing techniques. Of particular interest in the healthcare domain are adverse drug effects, which may be introduced in online posts, and can be effectively centralized and investigated. This paper presents our Multi-Task Learning architecture using pretrained Transformer-based language models employed for the Social Media Mining for Health Applications Shared Task 2021, where we tackle the three subtasks of Task 1, namely: classification of tweets containing adverse effects (subtask 1a), extraction of text spans containing adverse effects (subtask 1b), and adverse effects resolution (subtask 1c). Our best performing model ranked first on the test set at subtask 1b with an F_1-*score* of 51% (P = 51%; R = 51%). Promising results were obtained on subtask 1a (F_1-*score* = 44%; P = 45%; R = 43%), whereas subtask 1c was by far the most difficult task and an F_1-*score* of only 17% (P = 17%; R = 18%) was obtained.

1 Introduction

Information extraction from social media is widely studied nowadays, as platforms like Facebook, Twitter, Instagram, or Reddit become the main place for people to share their opinions and experiences. Concurrently, a wide range of applications with completely different topics arises with the current advances in Natural Language Processing (NLP), as the volume of posted information has become impossible to be manually analysed. For example, the Social Media Mining for Health (SMM4H) Applications Shared Task (Sarker and Gonzalez-Hernandez, 2017) is focused on health applications and introduces a dataset of annotated tweets with the aim to analyse adverse drug effects (ADE) mentioned by users. This year's edition (Magge et al., 2021) proposed eight different tasks, out of which we focused on the first task entitled *Classification, Extraction and Normalization of Adverse Effect mentions in English tweets*. This task was further divided into three subtasks, as follows. *Subtask 1a* was a binary classification task of tweets, focused on identifying whether the message contains ADE or not. *Subtask 1b* was a named entity recognition task on top of subtask 1a, centered on extracting the span of text containing the ADE of the medication. *Subtask 1c* was a named entity resolution task on top of both previous subtasks, aimed at predicting the normalized concept of the extracted adverse effect from the preferred terms included in the Medical Dictionary for Regulatory Activities (MedDRA)[1].

All three subtasks are addressed simultaneously using a Multi-Task Learning (MTL) architecture (Caruana, 1997) that leverages acquired knowledge from one subtask to another. Furthermore, we approached the challenge of unbalanced classes in the first subtask by considering class weights and by augmenting the training data set.

The paper is structured as follows. The second section describes previous work that inspired our solution, while the third section presents our employed method. The fourth section presents the results of our work, followed by discussions and the final section that summarizes our findings and presents future research paths.

2 Related Work

2.1 Health-Related Applications

Given that Task 1 was present in previous editions of the SMM4h shared task (Weissenbacher et al., 2018, 2019; Klein et al., 2020), several approaches were employed to address its challenges. For example, the winning team from 2019 (Miftahutdinov

[1] https://www.meddra.org/

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 44–51
June 10, 2021. ©2021 Association for Computational Linguistics

et al., 2019) used an ensemble of BioBERT-CRF models for the ADE extraction task, while addressing the resolution task as a classification. The system proposed by Miftahutdinov et al. (2020) ranked first at the end-to-end 2020 competition using the pretrained EnDR-BERT (Tutubalina et al., 2020) and the CSIRO Adverse Drug Event Corpus (CADEC) (Karimi et al., 2015) for further training the model. In addition, Dima et al. (2020) showed that bidirectional Transformers trained using class weighting, together with ensembles that combine various configurations, achieve an F1-score of .705 on the dataset made available for that edition of the competition.

2.2 MTL-Based Methods

Multi-Task Learning represents a training strategy where a shared model is simultaneously learning multiple tasks. Ruder (2017) analysed the techniques applied in MTL and compared the *hard parameter sharing* and *soft parameter sharing* paradigms, concluding that the former is still pervasive in nowadays approaches. MTL proved to fasten the convergence and to improve the model performance in a variety of NLP applications, including named entity recognition (Aguilar et al., 2018), fake news detection (Wu et al., 2019), multilingual offensive language identification (Chen et al., 2020b), sentiment analysis (Zaharia et al., 2020), humor classification (Vlad et al., 2020), recommender systems (Tang et al., 2020), and even question answering (Kongyoung et al., 2020). MTL also increases performance in conjunction with semi-supervised learning (Liu et al., 2007), curriculum learning (Dong et al., 2017), sequence-to-sequence (Zaremoodi and Haffari, 2018), reinforcement learning (Gupta et al., 2020), and adversarial learning (Liu et al., 2017).

3 Method

3.1 Corpus

The SMM4H 2021 Task 1 dataset included 17,385 training samples out of which 1,235 (7.10%) belong to the positive class (i.e., contain ADE), as well as 915 samples in the development set out of which 65 are labeled as positive; hence, a challenge consists of the unbalanced distribution of the two classes.

Subtask 1c required labeling the extracted text span with the corresponding MedDRA term; the number of possible labels exceeds 23,000. Only 476 labels are present in the training set, denoting that most labels are not covered at all. Additionally, the number of appearances of each ID has a long-tail distribution (see Figure 1), with some IDs being present in more than 60 examples and most IDs occurring in less than 4 examples.

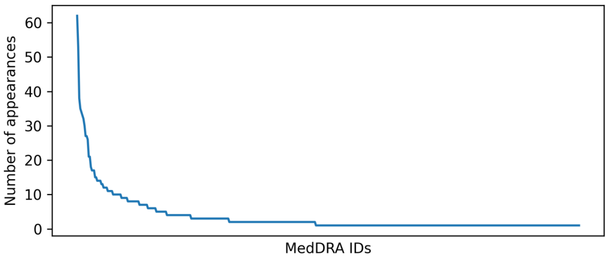

Figure 1: Histogram of MedDRA IDs present in the training set.

3.2 Multi-Task Learning Neural Architecture

A MTL architecture based on hard parameter sharing (Ruder, 2017) was employed for Task 1 (see Figure 2). Given that all three subtasks are highly related, our assumption was that knowledge acquired while learning one subtask would help in increasing performance on the other two. Three modules are added on top of BioBERT (Lee et al., 2020): (a) the *Classifier* - a binary classifier for tweet classification in subtask 1a, b) the *Extractor* - a named entity recognition layer for ADE span extraction in subtask 1b, and c) the *Normalizer* - a multi-class classifier for span resolution in subtask 1c. All three modules share the same pre-trained BERT encoder; the first 11 layers out of 12 were frozen, whereas the last layer was kept as a shared trainable encoder.

The training dataset was processed in the following manner. The positive tweets from the training set were selected for subtask 1b, and each token was tagged with either "O" (outside adverse effect) or "AE" (adverse effect entity). Two approaches were considered for subtask 1c: (a) create a dataset using the spans labeled with their corresponding PTID, and (b) concatenate the span tokens with the corresponding tweets as: *[CLS] <ADE span > [SEP] <entire tweet> [SEP]*. The second approach aimed to leverage context information in the MedDRA ID prediction.

The modules for the three subtasks were trained in parallel. At each training step, a batch from the training datasets was randomly chosen with the

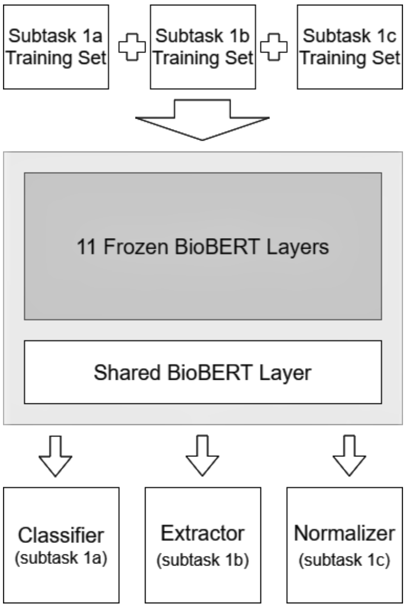

Figure 2: The overall architecture of the proposed multi-task framework.

following probability:

$$p_i = \frac{\frac{size(D_i)}{size(b_i)}}{\left(\sum_{k=1}^{n} \frac{size(D_k)}{size(b_k)}\right)} \tag{1}$$

where D_i represents the dataset for subtask i, and b_i represents a mini-batch from D_i.

All three subtasks minimize cross-entropy loss (see Equation 2), whereas only subtask 1a considers values different from one for weights w_j:

$$L_{Task} = -w_j \sum_i (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)) \tag{2}$$

The final loss minimized at each step k of Algorithm 1 is expressed in Equation 3:

$$L_k = t_k^1 L_{cls} + t_k^2 L_{ner} + t_k^3 L_{norm}, \tag{3}$$

where $t_k^i = 1$ when task i is in training, or $t_k^i = 0$ otherwise.

Algorithm 2 describes the processing pipeline which begins by passing the input tweet through the *Classifier*. If a tweet is labeled as not containing an adverse effect, the label is memorized and the flow stops. Otherwise, the tweet is passed to the *Extractor* and, afterwards, to the *Normalizer*. Line 10 highlights that the input tweet can also be used besides the text span that contains the ADE, in order to leverage the context information in predicting the MedDRA ID; this feature is optional and can be deactivated in certain configurations. Moreover, the feedback loop from the *Extractor* considering the label of the tweet (line 12) is optional.

Algorithm 1: Multi-task training algorithm

```
Initialize model parameters $\Theta$ from
  BioBERT using transfer learning;
Compute probabilities $p_i$ for each subtask $i$
  using Equation 1;
Shuffle and pack datasets into mini-batches
  $D_1, D_2, D_3$;
for $N$ training steps do
    Choose task $i$ with probability $p_i$;
    if $D_i$ is empty then
        Shuffle and re-pack $D_i$;
    end
    Randomly choose mini-batch $b$ from $D_i$;
    Compute task specific loss $L_i$ on $b$;
    Update $\Theta$ using $L_i$;
end
```

Algorithm 2: Task 1 prediction algorithm

```
1  Initialize $Classifier$;
2  Initialize $Extractor$;
3  Initialize $Normalizer$;
4  Load dataset $D$;
5  for $tweet$ in $D$ do
6      $label \leftarrow Classifier(tweet)$;
7      if $label$ is $ADE$ then
8          $S \leftarrow Extractor(tweet)$;
9          if $S$ is not empty then
10             $I \leftarrow Normalizer(span,$
                 $tweet)$ for each $span$ in $S$;
11         else
12             Change $label$ to $NoADE$;
13         end
14     end
15     Save $(label, S, I)$;
16 end
```

3.3 Implementation Details

Language Models: We experimented with BERT-base (Devlin et al., 2019) and with the domain-specific Transformers, namely BioBERT and BioClinicalBERT (Alsentzer et al., 2019). After a preliminary fine-tuning on the subtask 1a, the most promising results were obtained by BioBERT.

Given the limited resources available, we kept it as the default pre-trained solution in all further experiments.

Hyperparameters: All three modules (*Classifier*, *Extractor*, and *Normalizer*) were trained with a learning rate of $5e-5$. Batch sizes of 64 were used for subtask 1a that had most entries, while batch sizes of 16 were considered for subtasks 1b, and 1c in which only positive samples are considered. Training was performed for 30 epochs, computing the performance on the validation set after each epoch and saving the system that performed best.

Class Weights: The class unbalance problem from subtask 1a is addressed using the weighted version of the cross-entropy loss. The weights of the two classes were computed using the *balanced* heuristic (King and Zeng, 2001) from the scikit-learn library (Pedregosa et al., 2011).

Augmented Training Dataset: Another explored solution for the unbalance in subtask 1a consists in augmenting the poorly represented class (the positive class). We leverage the predefined augmentation approaches integrated into the TextAttack library (Morris et al., 2020). New positive examples are generated by char swapping, by replacing words with synonyms from the WordNet thesaurus (Miller, 1995), and by using methods from the CheckList testing - i.e., transformations like location replacement or number alteration (Ribeiro et al., 2020). Five positive examples are automatically added for each initial positive sample, thus increasing the proportion of the poorly represented class from 7% to almost 45%.

Class Number Reduction for the Normalizer: We considered subtask 1c a multi-class classification task where the *Normalizer* module receives as input the text span containing an ADE (i.e., the output of the *Extractor* module) and classifies the span into one of the classes (i.e., MedDRA PTIDs) present in the training set. The distribution of the 476 MedDRA IDs influenced us to reduce the number of classes. As such, the final classifier considers only the most frequent 108 PTIDs (i.e., IDs that appear more than three times in the training dataset). There were too few examples to properly generalize for all PTIDs; however, the module covers only 69.5% of the training samples.

4 Results

Four configurations were compared in terms of performance. The first configuration (*MTL*) is a baseline relying on the previously described MTL architecture. Weighted binary cross-entropy loss and feedback from the *Extractor* to the *Normalizer* (Line 12 from Algorithm 2) are enabled, but the *Normalizer* uses only the ADE span, without the entire tweet (Line 10 from Algorithm 2).

The second configuration (*MTL + BoostingEnsemble*) starts from *MTL*, but instead of the simple *Classifier* model, it uses an ensemble of three models trained in a boosting manner. The first classifier (*Classifier1*) is identical to the classifier from the first configuration. The second classifier (*Classifier2*) was trained on a modified training set in which the miss-classifies examples from *Classifier1* are over-sampled by a factor of three, whereas the correctly classified examples are down-sampled by the same factor. The third classifier, *Classifier3*, is also trained on a modified training set in which examples with different results from *Classifier1* and *Classifier2* are over-sampled by a factor of three, while the rest of the examples are down-sampled by the same factor.

The third configuration, denoted *MTL + EnhancedEnsemble*, further tries to improve the performance of the second configuration by adding two more classifiers to the ensemble, *Classifier4* and *Classifier5*, trained now on the augmented training set while considering equivalent over- and down-sampling approaches.

The fourth configuration, namely *MTL + EnhancedNormalizer*, is similar to the first configuration, but with the *Normalizer* is trained on both the ADE span and the entire tweet.

Table 1 introduces the comparative results for all configurations. While considering the development dataset, *MTL + EnhancedEnsemble* obtains the best performance for subtasks 1a and 1b, while *MTL + EnhancedNormalizer* has the highest *F1-score* for subtask 1c. In terms of the test dataset, *MTL* has the highest F1-score for subtask 1a - although the other configurations gain a boost in precision, recall is negatively influenced; *MTL + BoostingEnsemble* has the best performance on subtask 1b, whereas *MTL + EnhancedNormalizer* remains the best configuration for subtask 1c. Although *MTL + EnhancedEnsemble* has better results while integrating the augmented dataset, there are no improvements on the test dataset.

Model	**Subtask**	**Development**			**Test**		
		P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
MTL	1a	58.9	66.1	62.3	45.2	43.4	**44.3**
	1b	44.3	64.1	52.4	44.2	53.6	48.5
	1c	14.2	21.8	17.2	14.5	18.7	16.4
MTL + BoostingEnsemble*	1a	61.7	64.6	63.1	49.1	39.3	43.7
	1b	44.1	61.9	51.5	51.4	51.4	**51.4**
	1c	15.5	22.9	18.5	16.0	17.0	16.0
MTL + EnhancedEnsemble*	1a	65.1	62.1	**63.5**	48.8	36.6	42.0
	1b	50.8	62.3	**56.0**	51.4	49.1	50.0
	1c	15.7	20.6	17.9	15.8	16.0	16.0
MTL + EnhancedNormalizer	1c	17.2	24.1	**20.0**	16.9	17.9	**17.4**

Table 1: Evaluation of configurations for each subtask of SMM4H Task 1.
* marks the official submissions.

5 Discussions

Table 2 introduces classification problems that provide additional insights on how our *MTL + BoostingEnsemble* model works. Overall, it correctly extracts and classifies most text spans containing usual words for adverse effects (e.g., "sick") but, it has occasional difficulties in distinguishing between the desired effect of a medication and its adverse effects. For instance, in the first example from Table 2, our model does not make the association that the described medication is supposed to help the subject sleep, but, in contrast, it assumes sleepiness as an adverse effect.

Another limitation of our method is highlighted in the second example. The MedDRA term of *Slurred speech* is a rather rare label, not even present in the training set. Even though our system correctly extracts the span containing the adverse effect, it is unable to correctly predict the *ptid*.

The false positive example of "drunk" labeled as *Drunk like effect* shows that our model finds it hard to discern appearances from facts. A similar bias can be observed in the third example, where the model fails to extract the spans "sleep" and "stomach is a cement mixer" most likely because it learned that interrogations ask about adverse effects rather than offer information about them.

The fourth example denotes subtle errors, like grasping the difference between the MedDRA terms of *Sleepiness* and *Somnolence*, which are likely to be mislabeled even by humans.

While considering the differences between development and test set performances, another limi-

Annotated sample		**Model prediction**	
	MedDRA		MedDRA
...trazodone, it takes the light right outta your eyes...		...trazodone, it takes the light right outta your eyes ...	*Sleepiness*
one of the things i hate most about quetiapine is when i take it for the first few hours i slur my words, so people assume i'm merely drunk.	*Slurred speech*	one of the things i hate most about quetiapine is when i take it for the first few hours i slur my words, so people assume i'm merely drunk .	*Fluid retention*, *Drunk-like effect*
ciprofloxacin: how do you expect to sleep when your stomach is a cement mixer ?	*Sleeplessness*, *Stomach perforation*	ciprofloxacin: how do you expect to sleep when your stomach is a cement mixer?	
just woke up. since i started on the higher dose of quetiapine i'm sleeping even more ...; i feel knackered when i wake .	*Sleepiness*, *Groggy on awakening*	just woke up. since i started on the higher dose of quetiapine i'm sleeping even more ...; i feel knackered when i wake.	*Somnolence*, *Feeling stoned*

Table 2: Examples from the validation set obtained using *MTL + BoostingEnsemble*. Note that the MedDRA IDs were replaced by their *Preferred Terms*.

tation emerges, namely that our configurations did not generalized as expected on the test set for subtask 1a. This is argued by the reduced development set which contains only 5% of the provided labeled examples, coupled with our training procedure of always saving the model at its best validation score.

6 Conclusions and Future Work

We introduced a Transformer-based Multi-Task Learning architecture employed for Task 1 from the Social Media Mining for Health Applications Shared Task 2021. Task 1 was concerned with the classification of tweets incorporating adverse effects of medication and, for the positive tweets, with the extraction and normalization of the adverse effects. We started from a pretrained domain-specific BERT language model (i.e., BioBERT) which was further finetuned in a multi-task setting. A hard parameter sharing MTL model was trained on the three subtasks of SMM4H Task 1. Furthermore, class weights and data augmentation were considered to overcome the problem of the unbalanced dataset from subtask 1a.

Our model achieved the highest score for subtask 1b (i.e., adverse effect span detection) with an F_1-score of 51%, arguing that MTL can enhance adverse effect extraction from social media posts. In terms of future work, adversarial training (Miyato et al., 2018; Chen et al., 2020a) will be considered to improve the robustness of our approach.

References

Gustavo Aguilar, Adrian Pastor López-Monroy, Fabio A González, and Thamar Solorio. 2018. Modeling noisiness to recognize named entities using multitask neural networks on social media. 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 1401–1412.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Kejiang Chen, Yuefeng Chen, Hang Zhou, Xiaofeng Mao, Yuhong Li, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2020a. Self-supervised adversarial training. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2218–2222. IEEE.

Po Chun Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020b. Ntu_nlp at semeval-2020 task 12: Identifying offensive tweets using hierarchical multi-task learning approach. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2105–2110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

George-Andrei Dima, Andrei-Marius Avram, and Dumitru-Clementin Cercel. 2020. Approaching smm4h 2020 with ensembles of bert flavours. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 153–157.

Qi Dong, Shaogang Gong, and Xiatian Zhu. 2017. Multi-task curriculum transfer deep learning of clothing attributes. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 520–529. IEEE.

Deepak Gupta, Hardik Chauhan, Ravi Tej Akella, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Reinforced multi-task approach for multi-hop question generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2760–2775.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#smm4h) shared tasks at coling 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task.*

Sarawoot Kongyoung, Craig Macdonald, and Iadh Ounis. 2020. Multi-task learning using dynamic task weighting for conversational question answering. In *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*, pages 17–26.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Pengfei Liu, Xipeng Qiu, and Xuan-Jing 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)*, pages 1–10.

Qiuhua Liu, Xuejun Liao, and Lawrence Carin. 2007. Semi-supervised multitask learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 937–944.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2019. Kfu nlp team at smm4h 2019 tasks: Want to extract adverse drugs reactions from tweets? bert to the rescue. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 52–57.

Zulfat Miftahutdinov, Andrey Sakhovskiy, and Elena Tutubalina. 2020. Kfu nlp team at smm4h 2020 tasks: Cross-lingual transfer learning with pretrained language models for drug reactions. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 51–56.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

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

Abeed Sarker and Graciela Gonzalez-Hernandez. 2017. Overview of the second social media mining for health (smm4h) shared tasks at amia 2017. *Training*, 1(10,822):1239.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Fourteenth ACM Conference on Recommender Systems*, pages 269–278.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2020. The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*.

George-Alexandru Vlad, George-Eduard Zaharia, Dumitru-Clementin Cercel, Costin Chiru, and Stefan Trausan-Matu. 2020. Upb at semeval-2020 task 8: Joint textual and visual modeling in a multitask learning architecture for memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1208–1214.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 21–30.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, pages 13–16.

Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. 2019. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4636–4645.

George-Eduard Zaharia, George-Alexandru Vlad, Dumitru-Clementin Cercel, Traian Rebedea, and Costin Chiru. 2020. Upb at semeval-2020 task 9: Identifying sentiment in code-mixed social media texts using transformers and multi-task learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1322–1330.

Poorya Zaremoodi and Gholamreza Haffari. 2018. Neural machine translation for bilingually scarce scenarios: a deep multi-task learning approach. 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 1356–1365.