MIDAS@SMM4H-2019: Identifying Adverse Drug Reactions and Personal Health Experience Mentions from Twitter

Sarthak Anand,[3] **Debanjan Mahata,**[1] **Haimin Zhang,**[1] **Simra Shahid,**[2]
Laiba Mehnaz,[2] **Yaman Kumar**,[5] **Rajiv Ratn Shah**,[4]

[1]Bloomberg, USA, [2]DTU-Delhi, India, [3]NSIT-Delhi, India, [4]IIIT-Delhi, India, [5]Adobe, India,
sarthaka.ic@nsit.net.in, dmahata@bloomberg.net, hzhang449@bloomberg.net,
simrashahid_bt2k16@dtu.ac.in, laibamehnaz@dtu.ac.in, ykumar@adobe.com,
rajivratn@iiitd.ac.in

Abstract

In this paper, we present our approach and the system description for the Social Media Mining for Health Applications (SMM4H) Shared Task 1,2 and 4 (2019). Our main contribution is to show the effectiveness of Transfer Learning approaches like BERT and ULMFiT, and how they generalize for the classification tasks like *identification of adverse drug reaction mentions* and *reporting of personal health problems* in tweets. We show the use of stacked embeddings combined with BLSTM+CRF tagger for identifying spans mentioning adverse drug reactions in tweets. We also show that these approaches perform well even with imbalanced dataset in comparison to undersampling and oversampling.

1 Introduction

Drugs administered for alleviating common sufferings are the fourth biggest cause of death in US, following cancer and heart diseases (Giacomini et al., 2007), making it one of the most important medical problems for the human society. While heart diseases and cancer are commonly reported and studied, adverse reactions to drugs either goes unreported or is confused or lost within other narratives. While it is the onus of the government and the society as a whole to tackle the first task, the second one is an overwhelmingly computational task.

With the advent of universal internet and smartphones, reportrage of incidents is generally increasing, thanks to a host of social media platforms like *Twitter, Facebook, Instagram, etc*. Hence, this unique situation presents a challenging as well as rewarding opportunity to improve our current computational systems for dealing with the existing incidents more sensibly and increase their reportage with the use of electronic media.

With this motivation, four shared tasks were conducted as part of *Social Media Mining for Health Applications (SMM4H) Workshop 2019* (Weissenbacher et al., 2019). Our team participated in Tasks - 1, 2 and 4 of the workshop. The problems for these tasks were:

Problem Definition Sub-task 1: *Given a labeled dataset D of tweets, the objective of the task is to learn a classification/prediction function that can predict a label l for a given tweet t, where $l \in$ {reporting adverse effects of drugs (ADR) - 1, no adverse effects of drugs (non-ADR) - 0}.*

Example of tweets mentioning adverse drug reactions:

• *I feel siiiiiiiiiiiiick. Damn you venlafaxine.*

• *Who need alcohol when you have gabapentin and tramadol that makes you feel drunk at 12oclock.*

Problem Definition Sub-task 2: The motive of this sub-task is to first discern ADR tweets from the non-ADR ones and then identify the span of a tweet where an adverse drug effect is reported.

An example of a span from a tweet that represents the mention of adverse drug reactions:

• *losing it. could not remember the word power strip. wonder which drug is doing this memory lapse thing. my guess the cymbalta. #helps*, where *not remember* is the adverse drug reaction that needs to be identified and extracted from the tweet, which is most likely caused by the intake of the drug named *cymbalta*.

Problem Definition Sub-task 4: *Given a labeled dataset D of tweets, the objective of the task is to learn a classification/prediction function that can predict a label l for a given tweet t, where $l \in$ {reporting personal health experience - 1, no mention of personal health experience - 0}.*

Example of tweets reporting personal health experience mentions:

• *This flu shot got my arm killing me.*

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 127–132
Florence, Italy, August 2, 2019.

• *man i am so sick i feel terrible i got all the symptoms of the swine flu i am scared.*

Our Contributions: Towards the objectives of the tasks as described above, we present some of our contributions in this paper:

1. We train ULMFit and BERT models for Tasks 1 and 4, and show that these models are agnostic to the effects of undersampling and oversampling, given a highly imbalanced dataset.

2. We make an initial attempt in studying the effectiveness of transfer learning using ULMFit and BERT for the problems in the domain of health care pertaining to the shared tasks.

3. We show the use of stacked embeddings combined with BLSTM+CRF tagger for identifying spans mentioning adverse drug reactions in tweets.

4. We also show the use of combining pretrained BERT embeddings with Glove embeddings fed to a BLSTM text classifier for sub-task-1 and sub-task-4.

2 Related Work

In general, self reporting of drug effects by patients is a highly noisy source of data. However, even after being noisy, it captures quite a lot of information which might not be available in other cleaner sources of data such as limited clinical trials or a doctor's office (Leaman et al., 2010). Taking cognizance of this, the International Society of Drug Bulletins in 2005 said, "...patient reporting systems should periodically sample the scattered drug experiences patients reported on the internet...". This is an upcoming branch which lies at the intersection of information systems and medicine - *pharmacovigilance* (Leaman et al., 2010). Detecting and tracking information about certain diseases has been the focus of quite a lot of work (Nakhasi et al., 2012; Paul and Dredze, 2011). For instance, cancer investigation (Ofran et al., 2012), flu (Aramaki et al., 2011; Lamb et al., 2013) and depression (De Choudhury et al., 2013; Yazdavar et al., 2017). There has been some work in the domain of pharmacovigilance (Mahata et al., 2018b,a,c; Mathur et al., 2018; Sarker et al., 2018), recently as well.

The body of works most relevant to ours is the one which uses transfer learning on health domain.

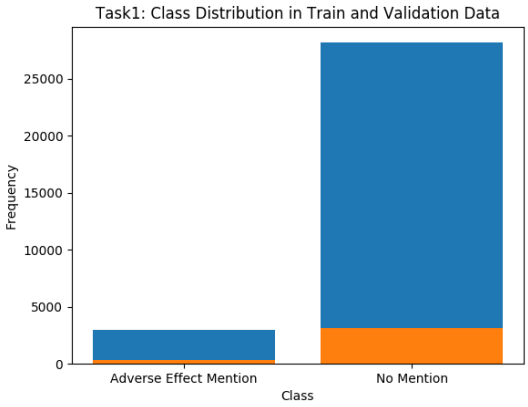

Figure 1: Distribution of classes in Train and Validation datasets for Sub-Task-1 (Identifying ADR and non-ADR tweets)

Normally, data in health domain is harder to get and process. Thus, many researchers have resorted to using transfer learning in order to deal with the data paucity. The works using transfer learning generally use word embeddings in order to improve the generalization of classification to unseen textual cases. In the context of this work we heavily use ULMFit (Howard and Ruder, 2018) and BERT (Devlin et al., 2018) for our experiments and make an initial attempt on how transfer learning in the domain of health works using them for the different text classification tasks of Social Media Mining for Health Workshop. Next, we give a brief description of the datasets used in this work for the different tasks.

3 Dataset

The dataset for the shared tasks was collected from the social networking website, *Twitter*. It consists of mentions of drug effects and other health related issues.

1. For the shared task 1, a total of 25,672 tweets are made available for training, out of which 2,374 contain adverse drug reaction (ADR) mention and the rest (23,298) do not. Only training data was provided by the organizers. For performing our experiments we segmented the provided dataset into train and validation splits. Figure 1 shows the distribution of data in the training and validation splits. The evaluation metric for this task was the F-score for the ADR class. Due to appreciable data bias, for the various experiments for this subtask, we oversample ADR tweets

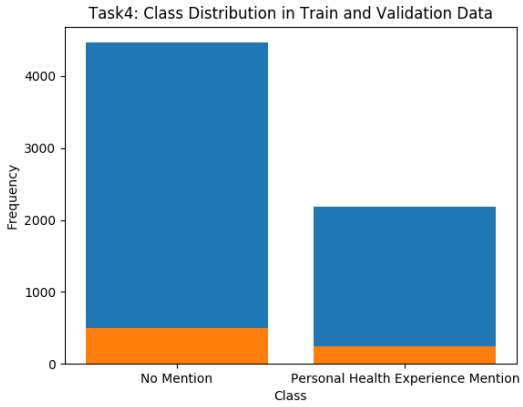

Figure 2: Distribution of classes in Train and Validation datasets for Sub-Task-4 (Identifying reporting of personal health experience mentions and no mentions in tweets)

and undersample non-ADR tweets. For oversampling, we just copy the ADR tweets and for undersampling, we randomly select a set of tweets such that the total number of tweets in both the sets becomes equal. For instance *"feeling a little dizzy from the quetiapine i just popped!"* represents a positive sample from the dataset while *"don't say no to pills! latuda won't kill!"* is a non-ADR tweet. We also try imbalanced proportions such as from 1:2 to 1:10 as well.

2. For the shared task 2, we got a total of 2,367 tweets out of which 1,212 were positive and 1,155 were negative. In the positive samples, the ADR portion was marked. For instance, the tweet *"friends! anybody taken #cipro? (antibiotic) complications?? big side effect is tendon rupture...figured my dr would know better?"* is an ADR tweet and the portion *"tendon rupture"* is where the author of the tweet mentions about ADR.

3. For the shared task 4, we were given a total of 10,876 tweets out of which only 7,388(67.9%) of the tweets were available on twitter for downloading. A total of 3,598 were positive and the rest were negative in original data. The positive tweets in this case contained a personal mention of ones health (for example, sharing health status or opinion) where as negative samples contained a generic discussion of the health issue, or some unrelated mention of the word. For instance, 9,832 is an example of tweet which contains flu-vaccination context in original data. Similarly, in the tweet 1,046, the author tries to discuss disease context of flu. For the available data we had 2,426 positive combined and 4,962 negative samples where the author is initiating general health discussion as opposed to mentioning any particular context of flu. For performing our experiments we segmented the provided dataset into train and validation splits. Figure 2 shows the distribution of data in the training and validation splits.

3.1 Preprocessing

Before feeding the dataset to any machine learning model we took some steps to process the data. We point to those steps in this section. Normalization of tokens were done using some hand-crafted rules mainly for dealing with short forms such as *thru*(through), *abt*(about), etc. The *'@user'* and URL tokens were removed. The hashtags that contained two or more words were segmented into their component words using *ekphrasis* library[1]. For example *#NotFeelingWell* was converted to not feeling well.

3.2 Training Models

For all the tasks, we mainly concentrated in training recently introduced ULMFit and BERT models that are well known for their transfer learning capabilities and generalizing well for various natural language processing tasks across different domains. We describe our models in this section. We extensively used fast.ai[2], bert[3], and flair[4] for training our models related to all the tasks. The different models trained and their corresponding hyperparamaters chosen for the tasks are presented in Table 1. We provide their brief description next.

ULMFiT- We used ULMFit (Howard and Ruder, 2018) for tasks 1 and 4. One of the main advantages of training ULMFiT is that it works very well for a small dataset as provided in the task and also avoids the process of training a classification model from scratch. This avoids overfitting. We have used the base (*fast.ai*) implementation of this model.

The ULMFiT model has mainly two parts, the language model and the classification model.

[1]https://github.com/cbaziotis/ekphrasis
[2]http://nlp.fast.ai/category/classification.html
[3]https://github.com/google-research/bert
[4]https://github.com/zalandoresearch/flair

Tasks	Models	Hyperparameters
Task 1 (Identification of Tweets mentioining ADR)	BERT (Submission 1)	batch_size=32, learning_rate=2e-5, epochs=4
	ULMFit (Submission 2)	batch size=72, learning_rate= 3e-2, bptt=70, epochs= 8, embedding_size=400, hidden_size=1150, number_of_layers=3
	BLSTM (Submission 3)	Pretrained Embeddings - BERT + Twitter Glove learning_rate=0.1, mini_batch_size=32, anneal_factor=0.5, patience=5, max_epochs=50, lstm_units=512, dense-size=256
Task 2 (ADR span extraction from Tweets)	BLSTM + CRF (Submission 1)	Stacked Pretrained Embeddings - BERT+Twitter Glove hidden units=256, learning_rate=0.1, epochs=150, batch_size=32
	BLSTM + CRF (Submission 2)	Stacked Pretrained Embeddings - BERT+Twitter Glove + Flair hidden units=256, learning_rate=0.1, epochs=150, batch_size=32
Task 4 (Identification of Tweets reporting personal health experience)	BERT (Submission 1)	batch_size=32, learning_rate=2e-5, epochs=4
	BLSTM (Submission 2)	Pretrained Embeddings - BERT + Twitter Glove learning_rate=0.1, mini_batch_size=32, anneal_factor=0.5, patience=5, max_epochs=50, lstm_units=512, dense-size=256
	BLSTM (Submission 3)	Pretrained Embeddings - BERT + Flair learning_rate=0.1, mini_batch_size=32, anneal_factor=0.5, patience=5, max_epochs=50, lstm_units=512, dense-size=256

Table 1: Model architectures and their corresponding hyperparameters of all the submissions by team MIDAS for sub-tasks 1, 2 and 4.

We observe that fine-tuning the language model on a larger dataset provides a significant improvement in the performace (Tuhin Chakrabarty, 2019). Therefore, we fine-tune the language model over 1,90,823 tweets containing 250-drug related mentions (Sarker and Gonzalez, 2015). Default (*fast.ai*) parameters were used to train the language models. Finally, we find the best hyperparameters and train the classifier over the original training data.

BERT - We use the provided Tensorflow implementation of BERT and fine-tune BERT-base-uncased. We find the best parameters and train the model over original dataset.

BLSTM - We train a bidirectional LSTM text classifier and feed different types of pretrained embeddings as presented in the Table 1. It is important to note that due to the long time needed for training the BLSTM models with the embeddings and unavailability of GPUs, we could not finish the training before submitting our results for the test data provided by the organizers. We would like to make our predictions on the final model and keep it as a future work.

BLSTM+CRF Tagger - We treated the problem posed in sub-task 2 as a named entity extraction and recognition problem. The text span corresponding to an adverse drug reaction mention is treated as an entity, that further needs to be classified into one of the two categories ADR or non-ADR. Following the current state-of-the-art, we trained a BLSTM+CRF tagger implemented in the flair library (referred above). Apart from that, we also used the BLSTM+CRF architecture with two different combinations of stacked embeddings.

Next, we present the results obtained on the test data provided by the organizers for sub-tasks 1, 2 and 4.

4 Results

4.1 Task-1: Identifying Tweets Mentioning Adverse Drug Reactions

Model	F1	Precision	Recall
BERT	0.5759	0.5615	0.5911
ULMFiT	**0.5988**	0.6647	0.5447
BLSTM	0.5196	0.5891	0.4649

Table 2: Results for Task-1: Identifying Tweets Mentioning Adverse Drug Reactions

Table 2, presents the F1 scores on the test data for sub-task 1. The ULMFit model showed the best performance. As already mentioned, the data provided for sub-task 1 was highly imbal-

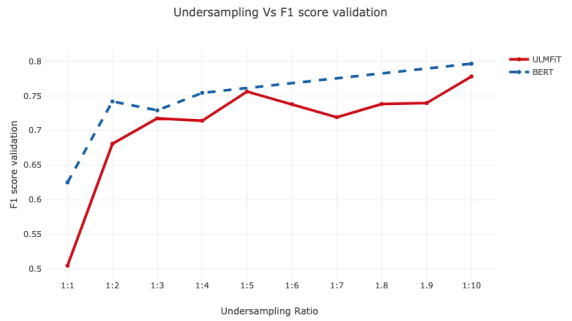

Figure 3: F1 score for ULMFit and BERT models trained on differently undersampled ratio of the classes (ADR : non-ADR).

anced. We performed undersampling with different ratios of the classes (ADR : non-ADR). Figure 3, presents the performance of ULMFit and BERT models on the training data for different undersampling ratios. We also tried oversampling, but didn't observe any improvement in performance. The best performance using both BERT and ULMFit was obtained without using any undersampling or oversampling. Therefore, the model that we used on the test data was trained on the full training dataset maintaining the given ratio of ADR:non-ADR tweets.

4.2 Task 2: Extraction of Adverse Effect Mentions

Model	Relaxed F1	Relaxed Precision	Relaxed Recall	Strict F1	Strict Precision	Strict Recall
BLSTM+CRF Tagger with Stacked Pretrained BERT and Twitter Glove Embeddings	0.638	0.532	0.796	0.315	0.262	0.395
BLSTM+CRF Tagger with Stacked Pretrained BERT and Flair Embeddings	**0.641**	0.537	0.793	**0.328**	0.274	0.409

Table 3: Results for Task-2: Extracting spans of text expressing adverse drug reactions in Tweets

Table 3, presents the performance scores for sub-task 2 on the test data. The different metrics as presented in the table were implemented by the organizers and the scores were provided by them.

4.3 Task 4: Generalized Identification of Personal Health Experience Mentions

The objective of the task is to classify whether a tweet contains a personal mention of ones health (for example, sharing ones own health status or opinion), as opposed to a more general discussion of the health issue, or an unrelated mention of the word. Each model was finally evaluated using four F1-scores - F1 for the held out influenza data, the second and third undisclosed context, and the F1-score overall. The results that our models obtained on the test data is presented in Table 4. As already mentioned that the BLSTM models trained using pretrained embeddings could not be completed. Inspite of the fully trained model, we do see a decent performance using BLSTM along with a combination of pretrained embeddings on the provided dataset.

Models	Accuracy	F1	Precision	Recall
Model 1 - BERT	0.8105	0.7453	0.9875	0.5985
Model 2 - BERT + Twitter Glove Embeddings	**0.8211**	**0.783**	0.8932	0.697
Model 3 - BERT + Flair Embeddings	0.8035	0.7544	0.8958	0.6515
Health Concerns Condition 1				
Model 1 - BERT	**0.9**	**0.8919**	1	0.8049
Model 2 - BERT + Twitter Glove Embeddings	0.8875	0.88	0.9706	0.8049
Model 3 - BERT + Flair Embeddings	0.8938	0.8859	0.9851	0.8049
Health Concerns Condition 2				
Model 1 - BERT	0.6377	0.359	0.875	0.2258
Model 2 - BERT + Twitter Glove Embeddings	**0.6667**	**0.5818**	0.6667	0.5161
Model 3 - BERT + Flair Embeddings	0.6087	0.4706	0.6	0.3871
Health Concerns Condition 3				
Model 1 - BERT	0.7679	0.48	1	0.3158
Model 2 - BERT + Twitter Glove Embeddings	**0.8214**	**0.6667**	0.9091	0.5263
Model 3 - BERT + Flair Embeddings	0.7857	0.5714	0.8889	0.4211

Table 4: Results for Task-4: Generalized identification of personal health experience mentions

5 Future Work and Conclusion

In this work, we presented our initial attempt to use BERT and ULMFit for text classification tasks related to the domain of pharmacovigilance. We obtained decent results for three different tasks organized as a shared task in Social Media Mining for Health Workshop - 2019. We noticed that the BERT and ULMFit were agnostic to undersampling and oversampling unlike previously observed performances on traditional text classifiers as reported on a similar task (Sarker et al., 2018), that was a part of the same workshop held in 2017. We consider our reported work in this paper as a preliminary attempt and would like to extend them in the future. As part of our future work we would like to train better models using BERT for all the three sub-tasks that we participated in, and would also like to interpret the predictions of the models. We think domain specific training of different embeddings could help and would like to try them in the future.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kathleen M Giacomini, Ronald M Krauss, Dan M Roden, Michel Eichelbaum, Michael R Hayden, and Yusuke Nakamura. 2007. When good drugs go bad. *Nature*, 446(7139):975.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

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

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing*, pages 117–125. Association for Computational Linguistics.

Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, and Jing Jiang. 2018a. Detecting personal intake of medicine from twitter. *IEEE Intelligent Systems*, 33(4):87–95.

Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, and Jing Jiang. 2018b. Did you take the pill?-detecting personal intake of medicine from twitter. *arXiv preprint arXiv:1808.02082*.

Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, et al. 2018c. # phramacovigilance-exploring deep learning techniques for identifying mentions of medication intake from twitter. *arXiv preprint arXiv:1805.06375*.

Puneet Mathur, Meghna Ayyar, Sahil Chopra, Simra Shahid, Laiba Mehnaz, and Rajiv Shah. 2018. Identification of emergency blood donation request on twitter. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–31.

Atul Nakhasi, Ralph Passarella, Sarah G Bell, Michael J Paul, Mark Dredze, and Peter Pronovost. 2012. Malpractice and malcontent: Analyzing medical complaints in twitter. In *2012 AAAI Fall Symposium Series*.

Yishai Ofran, Ora Paltiel, Dan Pelleg, Jacob M Rowe, and Elad Yom-Tov. 2012. Patterns of information-seeking for cancer on the internet: an analysis of real world data. *PloS one*, 7(9):e45921.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Smaranda Muresan Tuhin Chakrabarty. 2019. Columbianlp at semeval-2019 task 8: The answer is language model fine-tuning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1140–1144.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth Social Media Mining for Health (SMM4H) shared task at ACL 2019. In *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop Shared Task*.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198. ACM.