

# Transfer learning for health-related Twitter data

Anne Dirkson & Suzan Verberne

LIACS, Leiden University

Niels Bohrweg 1, Leiden, the Netherlands

{a.r.dirkson, s.verberne}@liacs.leidenuniv.nl

## Abstract

Transfer learning is promising for many NLP applications, especially in tasks with limited labeled data. This paper describes the methods developed by team TMRLeiden for the 2019 Social Media Mining for Health Applications (SMM4H) Shared Task. Our methods use state-of-the-art transfer learning methods to classify, extract and normalise adverse drug effects (ADRs) and to classify personal health mentions from health-related tweets. The code and fine-tuned models are publicly available.<sup>1</sup>

## 1 Introduction

Transfer learning is promising for NLP applications, as it enables the use of universal pre-trained language models (LMs) for domains that suffer from a shortage of annotated data or resources, such as health-related social media. Universal LMs have recently achieved state-of-the-art results on a range of NLP tasks, such as classification (Howard and Ruder, 2018) and named entity recognition (NER) (Akbik et al., 2018). For the Shared Task of the 2019 Social Media Mining for Health Applications (SMM4H) workshop team TMRLeiden focused on employing state-of-the-art transfer learning from universal LMs to investigate its potential in this domain.

## 2 Task descriptions

**ADR extraction** The purpose of **Subtask 1** (S1) is to classify tweets as containing an adverse drug response (ADR) or not. Subsequently, these ADR mentions are extracted in **Subtask 2** (S2) and normalized to MedDRA concept IDs in **Subtask 3** (S3). MedDRA (Medical Dictionary for Regulatory Activities) is an international, standardized medical terminology.<sup>2</sup>

<sup>1</sup><https://github.com/AnneDirkson/SharedTaskSMM4H2019>

<sup>2</sup><https://www.meddra.org/>

**Personal Health Mention Extraction** The goal of **Subtask 4** (S4) is to identify tweets that are personal health mentions, i.e. posts that mention a person who is affected as well as their specific condition (Karisani and Agichtein, 2018), as opposed to posts discussing health issues in general. Generalisability to both future data and different health domains is evaluated by including data from the same domain collected years after the training data, as well as data from entirely different disease domain.

## 3 Our approach

### 3.1 Preprocessing

We preprocessed all Twitter data using the lexical normalization pipeline by Sarker (2017). We also employed an in-house spelling correction method (Dirkson et al., 2019). Additionally, punctuation and non-UTF-8 characters were removed using regular expressions.

### 3.2 Additional Data

**Personal Health Mentions** For S4, the training data consists of data from one disease domain, namely influenza, in two contexts: having a flu infection and getting a flu vaccination. To improve generalisability, we supplemented this data with six labelled data sets from different disease domains (Karisani and Agichtein, 2018). We refer to this combined data set as S4+. For each subset, 10% was used for a combined validation set. For fine-tuning the ULMfit universal language model based on 28,595 Wikipedia articles (Wikitext-103) (Merity et al., 2017b), the DIEGO Drug Chat-ter corpus (Sarker and Gonzalez, 2017) was combined with the data from S1 and S4+ to form a larger unsupervised corpus of health-related Twitter data ('TwitterHealth'). For S4, fine-tuning was also attempted with only the S4+ data.

	S1	S2*	S3	S4	S4+
Dev	-	130	76	-	-
Train	14,634	910	1,756	6,996	11,832
Validation	1,626	130	76	777	1,314
Test	5000	1000	1000	TBA	TBA

Table 1: Data sets. \*Only tweets containing ADRs were used for developing the system. TBA: To be announced

**Concept Normalization** The MedDRA concept names and their aliases in both MedDRA and the Consumer Health Vocabulary<sup>3</sup> were used to supplement the data from S3. This data set is hereafter called S3+.

### 3.3 Text Classification

Text classification was performed with fast.ai ULMfit (Howard and Ruder, 2018). As recommended, the initial learning rate (LR) of 0.01 was determined manually by inspecting the log LR compared to the loss. Default language models were fine-tuned using AWD\_LSTM (Merity et al., 2017a) with (1) 1 cycle (LR = 0.01) for the last layer and then (2) 10 cycles (LR = 0.001) for all layers.

Subsequently, this model is used to train a classifier with  $F_1$  as the metric, a dropout of 0.5 and a momentum of (0.8,0.7), in line with the recommendations. Training is done with (1) 1 cycle (LR = 0.02) on the last layer; (2) unfreezing of the second-to-last layer; (3) another cycle running from a 10-fold decrease of the previous LR to this LR divided by 2.6<sup>4</sup> (as recommended in the fast.ai MOOC).<sup>4</sup> This is repeated for the next layer and then for all layers. The last step consists of multiple cycles until  $F_1$  starts to drop.

As an alternative classifier for S1, we used the absence of ADRs (noADE) according to the Bert embeddings NER method (see below) which was developed for the subsequent sub-task (S2) and aims to extract these ADR mentions. As a baseline for text classification, we used a Linear SVC with unigrams as features. The C parameter was tuned with a grid of 0.0001 to 1000 (steps of x10).

### 3.4 Named Entity Recognition

For S2, we experimented with different combinations of state-of-the-art Flair embeddings (Akbiik et al., 2018), classical Glove embeddings and

<sup>3</sup><https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/>

<sup>4</sup><https://course.fast.ai/>

Bert embeddings (Devlin et al., 2018) using the Flair package. We used pre-trained Flair embeddings based on a mix of Web data, Wikipedia and subtitles; and the ‘bert-base-uncased’ variant of Bert embeddings. We also experimented with Flair embeddings combined with Glove embeddings (dimensionality of 100) based on FastText embeddings trained on Wikipedia (GloveWiki) or on Twitter data (GloveTwitter). Training for all embeddings was done with initial LR of 0.1, batch size of 32 and max epochs set to 150.

As a baseline for NER, we used a CRF with the default L-BFGS training algorithm with Elastic Net regularization. As features for the CRF, we used the lowercased word, its suffix, the word shape and its POS tag.<sup>5</sup>

### 3.5 Concept normalization

For S3, pre-trained Glove embeddings were used to train document embeddings on the extracted ADR entities in the S3 data including or excluding the aliases from CHV (S3+) with concept IDs as labels. We used the default RNN in Flair with a hidden size of 512. Glove embeddings (dim = 100) were based on FastText embeddings trained on Wikipedia. Token embeddings were re-projected (dim = 256) before inputting to the RNN.

## 4 Results

	Method	$F_1$ (range)	P	R
Average*		0.502 (0.331)	0.535	0.505
Run1	ULMfit <sup>1</sup>	<b>0.533</b>	<b>0.642</b>	0.455
Run2	noADE	0.418	0.284	<b>0.792</b>

Table 2: Results for ADR Classification (S1). \*over all runs submitted <sup>1</sup>TwitterHealth data

For all four subtasks, our best transfer learning system consistently performs better than the average over all runs submitted to SMM4H. For classifying ADR mentions, our overall best performing system is a ULMfit model trained on the TwitterHealth corpus (see Table 2). Yet, the highest recall is attained by using the absence of named entities (noADE) as a classifier. This is in line with our validation results (see Table 6). For extracting ADRs, our best system is a combination of Bert with Flair embeddings without a separate classifier

<sup>5</sup><https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html>

	Method	relaxed F <sub>1</sub> (range)	relaxed P	relaxed R	strict F <sub>1</sub> (range)	strict P	strict R
	Average*	0.538 (0.486)	0.513	0.615	0.317 (0.422)	0.303	0.358
Run1	Bert+Flair <sup>+</sup>	<b>0.625</b>	0.555	<b>0.715</b>	<b>0.431</b>	0.381	<b>0.495</b>
Run2	Bert <sup>+</sup>	0.622	0.560	0.701	0.427	0.382	0.484
Run3	Bert+ADRClassifier	0.604	<b>0.718</b>	0.521	0.417	<b>0.494</b>	0.360

Table 3: Results for ADR Extraction(S2). \*over all runs submitted <sup>+</sup>No separate classifier for sentences containing ADRs

	Method	relaxed F <sub>1</sub> (range)	relaxed P	relaxed R	strict F <sub>1</sub> (range)	strict P	strict R
	Average*	0.297 (0.242)	0.291	0.312	0.212 (0.247)	0.205	0.224
Run1 <sup>+</sup>	RNN Docembeddings	<b>0.312</b>	<b>0.370</b>	0.270	<b>0.250</b>	<b>0.296</b>	0.216
Run2 <sup>+</sup>	RNN Docembeddings	0.303	0.272	0.343	0.244	0.218	0.277
Run3 <sup>+</sup>	RNN Docembeddings	0.302	0.267	<b>0.347</b>	0.246	0.218	<b>0.283</b>

Table 4: Results for concept normalization (S3). \*over all runs submitted <sup>+</sup>Runs same as S2 prior to concept normalization

	Method	Acc. (range)	F <sub>1</sub> (range)	P	R	
	Average*	0.781 (0.263)	0.701 (0.464)	0.902	0.585	
Run1	ULMfit with S4+ data	<i>Domain1</i>	0.869	0.859	0.952	0.781
		<i>Domain2</i>	0.638	0.419	0.750	0.290
		<i>Domain3</i>	0.786	0.539	1.000	0.368
		Mean	<b>0.793</b>	<b>0.726</b>	<b>0.940</b>	<b>0.591</b>
Run2	ULMfit with TwitterHealth data	<i>Domain1</i>	0.863	0.849	0.969	0.756
		<i>Domain2</i>	0.609	0.342	0.700	0.226
		<i>Domain3</i>	0.768	0.480	1.000	0.316
		Mean	0.786	0.716	0.928	0.583

Table 5: Results for personal health mention classification (S4). \*over all runs submitted

for sentences containing ADR mentions (see Table 3). However, using Bert embeddings alone *with* the ULMfit classifier from S1 appears to be more precise. During validation, we found that combinations of Glove embeddings (based on Twitter or Wikipedia) and Flair embeddings performed poorly compared to the submitted systems (see Table 7). For mapping the ADRs to MedDRA concepts, we only submitted one system with different preceding NER models (see Table 4), since adding the alias information (S3+) decreased both precision and recall (see Table 8). Our RNN document embeddings with only the S3 data, however, performed better than average. Lastly, for the classification of personal health mentions, our best classifier was a ULMfit model fine-tuned on the S4+ data (see Table 5), which outperformed the average result and the ULMfit model trained on the larger TwitterHealth corpus on all metrics. This system similarly outperformed the other ULMfit model on the validation data (see Table 9).

Method	F <sub>1</sub>	P	R
Baseline: Linear SVC (C=1.0)	0.475	0.526	0.433
ULMfit <sup>1</sup>	<b>0.574</b>	<b>0.574</b>	0.574
noADE	0.330	0.207	<b>0.823</b>

Table 6: Validation results for ADR classification (S1)  
<sup>1</sup>TwitterHealth data

Method	Micro-F <sub>1</sub>	P	R
Baseline: CRF	0.235	0.560	0.149
Flair+ GloveWiki	0.596	0.666	0.540
Flair+ GloveTwitter	0.577	0.655	0.515
Bert	0.640	<b>0.699</b>	0.590
Bert+Flair	<b>0.649</b>	<b>0.699</b>	<b>0.606</b>

Table 7: Validation results for ADR extraction (S2)

Method	F <sub>1</sub>	P	R
RNNDocembeddings with S3	<b>0.623</b>	<b>0.566</b>	<b>0.694</b>
RNNDocembeddings with S3+	0.253	0.171	0.482

Table 8: Validation results for concept normalization (S3)

Method	F <sub>1</sub>	P	R
Baseline: Linear SVC (C=0.1)	0.615	0.678	0.572
ULMfit with S4+ data	<b>0.712</b>	<b>0.743</b>	<b>0.701</b>
ULMfit with TwitterHealth data	0.692	0.738	0.676

Table 9: Mean validation results for personal health mention classification (S4) averaged over eight data sets of S4+

## 5 Conclusions

Transfer learning using default and recommended settings offers above average results for various NLP tasks using health-related Twitter data. More research is necessary to investigate whether state-of-the-art performance may be possible with further domain-specific adaptation, for instance by tuning hyper-parameters, training embeddings on medical data or by dealing with domain-specific vocabulary absent in the language model.

## References

- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. [Contextual string embeddings for sequence labeling](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). *ArXiv*.
- Anne Dirkson, Suzan Verberne, Gerard van Oortmerssen, Hans van Gelderblom, and Wessel Kraaij. 2019. [Lexical Normalization of User-Generated Medical Forum Data](#). In *Proceedings of 2019 ACL workshop Social Media Mining 4 Health (SMM4H)*.
- Jeremy Howard and Sebastian Ruder. 2018. [Universal language model fine-tuning for text classification](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Payam Karisani and Eugene Agichtein. 2018. [Did you really just have a heart attack?: Towards robust detection of personal health mentions in social media](#). In *Proceedings of the 2018 World Wide Web Conference, WWW '18*, pages 137–146, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017a. [Regularizing and Optimizing LSTM Language Models](#). *ArXiv*.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017b. [Pointer Sentinel Mixture Models](#). In *ICLR 2017*.
- Abeed Sarker. 2017. [A customizable pipeline for social media text normalization](#). *Social Network Analysis and Mining*, 7(1):45.
- Abeed Sarker and Graciela Gonzalez. 2017. [A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities](#). *Data in Brief*, 10:122–131.