SEED: SYMPTOM EXTRACTION FROM ENGLISH SOCIAL MEDIA POSTS USING DEEP LEARNING AND TRANSFER LEARNING

Arjun Magge
Perelman School of Medicine
University of Pennsylvania
Arjun.Magge@pennmedicine.upenn.edu

Karen O'Connor
Perelman School of Medicine
University of Pennsylvania
karoc@pennmedicine.upenn.edu

Matthew Scotch
School of Health Solutions
Arizona State University
Matthew.Scotch@asu.edu

Graciela Gonzalez-Hernandez
Perelman School of Medicine
University of Pennsylvania
gragon@pennmedicine.upenn.edu

February 10, 2021

ABSTRACT

The increase of social media usage across the globe has fueled efforts in public health research for mining valuable information such as medication use, adverse drug effects and reports of viral infections that directly and indirectly affect human health. Despite its significance, such information can be incredibly rare on social media. Mining such non-traditional sources for disease monitoring requires natural language processing techniques for extracting symptom mentions and normalizing them to standard terminologies for interpretability. In this work, we present the first version of a social media mining tool called SEED that detects symptom and disease mentions from social media posts such as Twitter and DailyStrength and further normalizes them into the UMLS terminology. Using multi-corpus training and deep learning models, the tool achieves an overall F1 score of 0.85 for extracting mentions of symptoms on a health forum dataset and an F1 score of 0.72 on a balanced Twitter dataset significantly improving over previously systems on the datasets. We apply the tool on recently collected Twitter posts that self-report COVID19 symptoms to observe if the SEED system can extract novel diseases and symptoms that were absent in the training data. By doing so, we describe the advantages and shortcomings of the tool and suggest techniques to overcome the limitations. The study results also draw attention to the potential of multi-corpus training for performance improvements and the need for continual training on newly obtained data for consistent performance amidst the ever-changing nature of the social media vocabulary.

Keywords Natural Language Processing · Deep Learning · Information Extraction · Social Media Mining · Pharmacovigilance

1 Introduction

Mining social media posts can reveal valuable information about early symptoms from emergent infectious diseases, medication use, abuse, and adherence, as well as adverse drug effects (ADEs) and environmental exposures that may have significant impact on human health. [1, 2, 3, 4] While the noisy nature of the data offers technical challenges for natural language processing methods, additional hurdles for data analysis such as selection bias in social media population and the relative scarcity of relevant postings could impede deriving signals for health policy and interventions. Despite these challenges, social media has been shown to be a promising complementary resource to established data sources. [5, 6, 7] While our prior work has addressed some of these challenges, we focus this effort on the identification of mentions of signs and symptoms of disease, an important building block for health monitoring efforts on medication safety, disease progression, or infectious disease spread.

This drug didn't end up helping at all. I tried it for my mood and anxiety but gave me vertigo. :(	mood (10024919: Low mood) anxiety (10002855: Anxiety) vertigo (10047340: Vertigo)
Is anybody else have terrible acne problems?	acne problems (10000496: Acne)
I can't see the point in taking -DRUG-. It's screwing up my sleep and making me depressed.	screwing up my sleep (10040995: Sleep disturbance) depressed (10012378: Depression)
It kinda did ok in keeping my anxiety under control. But now I need to get my heart rate down. Oh god.	anxiety (10002855: Anxiety) need to get my heart rate down (10019303: Heart rate increased)
Its dangerous. -DRUG- is supposed to treat depression and suicidal thoughts. I took it and got very depressed.	depression (10012378: Depression) suicidal thoughts (10042458: Suicidal ideation) depressed (10012378: Depression)

Figure 1: Examples illustrating mentions of symptoms across categories of indication (in blue) and ADEs (in red) in social media posts and assignment of normalization identifiers.

Although there has been previous work on extracting symptoms from Twitter, Facebook, Instagram, and more, [8, 9, 10] the work in this realm has been specific to particular contexts. For example, on detecting adverse drug events (ADEs), [**?**, 4] or symptoms specific to COVID-19. [11, 12, 13] To the best of our knowledge, the problem has not been addressed in general for all symptoms, regardless of the context in which they are mentioned, and no prior effort has addresses normalization.

In this work we present the first version of a social media mining tool that can detect generic disease symptom mentions. The tool also normalizes the detected symptoms into the UMLS (using MedDRA Preferred Term identifiers). These can be further mapped to other ontologies for phenotyping efforts. Figure 1 shows examples of mentions of symptoms in social media posts and their subcategories.

The primary objective of this work is to advance a system for extracting symptoms mentioned in social media (Twitter and health forums), regardless of the context in which they are mentioned. Following are the contributions of the work presented:

- We establish a new state-of-the-art performance for symptom extraction using a deep learning based NER.
- We present a symptom normalizer for converting the extracted spans to the expanded vocabulary from UMLS (using MedDRA Preferred Term identifiers).
- Considering no prior effort for generic symptom identification exists, we compare our system and establish state-of-the-art performance for ADE extraction and normalization.
- We make the extraction and normalization components publicly available (the dataset was already available) as part of the DRIP (DRug Insights for Pharmacovigilance) toolkit.

The remaining document is structured as follows: we discuss the dataset and models used for this work in the Materials and Methods section and discuss the evaluation results and its impact on research in the Results section.

2 Materials and Methods

2.1 Datasets

In this work, we reuse openly available annotated datasets for indications and ADEs. We use a total of three annotated datasets from two social media sources: Twitter (Tw-NER-v1 and Tw-Resolve) and DailyStrength (DS-NER). These datasets contain symptom mention annotations in two categories: ADE and Indication. Symptoms in the Twitter datasets have been annotated along with their drug spans and the DailyStrength dataset does not contain drug spans. By

Table 1: Summary of the datasets used for the experiments presented. As discussed before, we merge the NER datasets to detect symptoms by combining ADR and indication spans.

Corpus	Annotation Type	Training set	Test set	Positive posts
DailyStrength **(DS-NER)**	NER spans (ADR, Indication)	4720	1559	32%
Twitter **(Tw-NER-v1)**	NER spans (ADR, Indication)	1340	443	50%
Twitter **(Tw-Resolve)**	NER spans + MedDRA (ADR)	2276 [1]	1573	50%

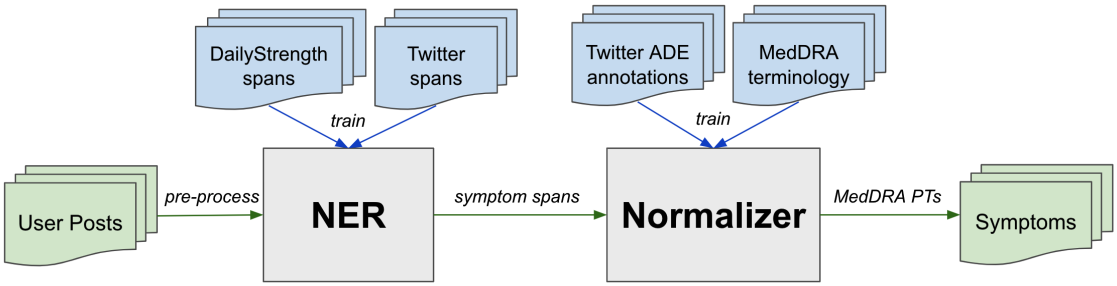

Figure 2: Training and inference components of the SEED tool used for extracting and normalizing symptom mentions in social media posts.

ignoring the drug spans and integrating both datasets, we make the system available to be run on user posts that do not necessarily contain drug mentions.

The above-mentioned datasets have been previously published for developing automated systems for detecting ADEs and indications on social media. [14] However, the Tw-Resolve dataset only contains span annotations for ADE and not indications, hence we only use this dataset for training the normalization system. [15] For each annotated ADE span in the dataset, it contains an annotated MedDRA identifier to the normalized medical term. We refer the readers to the original papers for details regarding data collection and annotation guidelines, and present a summary of the datasets used for experiments in this work in Table 1.

2.2 Concept Extraction - NER

For the NER tasks, we use the off-the-shelf deep learning-based Flair framework [16] to construct, train and test the models. The posts from both DailyStrength and Twitter training sets are tokenized using segtok and encoded into IOB2 format. 5% of the posts in the training set are separated into a development set for hyperparameter tuning and choosing the best model. NER models constructed in Flair contain the language representation layer followed by a bidirectional recurrent neural network (RNN) layer (dimension size 256) composed from gated recurrent units (GRU) which are then concatenated and passed through a fully connected layer. Outputs from the fully connected layer are then passed through a conditional random field (CRF) layer. The weights in the model are trained by the minimizing the loss computed using the negative log likelihood computed at the CRF layer.

The training was performed on a MacBook Pro 2019 with 8 cores and 16 gigabytes of RAM. The optimal settings were found to be a learning rate of 0.1 with a stochastic gradient descent (SGD) optimizer. The model was trained for 70 epochs and the model with the best performance on the development set was saved for testing its performance on the test sets. After experimenting with various forms of language representation techniques such as word2vec and Glove embeddings [17], FastText embeddings with enriched subword information [18] and BERT [19]. We found that FastText and BERT models performed significantly better than word2vec and Glove models.

2.3 Normalization

The normalization system is built on the off-the-shelf FastText classifier which employs multinomial logistic regression model with word and character n-gram features trained used a hierarchical softmax loss. [20] This normalization system setup provides faster training and inference capabilities compared to traditional classifiers that have scaling issues when the number of target labels are in the thousands. We find the hyperparameters are optimal when the hidden layer dimensions are set to 128, the word n-grams features are limited to maximum window of 3 and character n-grams features are limited to a maximum window of 5.

Table 2: Results of Multi-corpus Training on the Twitter and Dailystrength Datasets.

Test Set	**Twitter (ADR)**	**DailyStrength (ADR)**	**Twitter (Indication)**	**DailyStrength (Indication)**
Training Set	$P/R/F_1$	$P/R/F_1$	$P/R/F_1$	$P/R/F_1$
ADRMine [14]	0.76/0.68/0.72	0.86/0.78/0.82	-	-
Twitter	0.82/0.72/0.77	0.72/0.69/0.71	0.50/0.21/0.30	0.80/0.21/0.33
DailyStrength	0.77/0.57/0.66	**0.90**/0.82/0.86	0.19/**0.54**/0.28	0.84/**0.76/0.80**
Twitter + DailyStrength	**0.87/0.73/0.79**	0.89/**0.84/0.87**	**0.59**/0.46/**0.52**	**0.89**/0.71/0.79

The normalization system was trained on both the NER span texts in the training set as well as the MedDRA lower level terms (LLTs) such that predictions can be made and evaluated on the MedDRA preferred terms (PTs). Training on MedDRA LLTs in a self-supervised manner also allows for discovery of symptoms not annotated in the training set. We also train on additional expressions of symptoms gathered by integrating MedDRA PTs with terms from other ontologies in UMLS metathesaurus using their respective concept unique identifiers (CUI). [21]

3 Results and Discussion

The evaluation results for the NER are shown in Table 2. Here we see that multi-corpus training is highly beneficial for both NER datasets. Training on DailyStrength data increased the Twitter model's performance by 13 percentage points for ADRs and 23 percentage points for Indication extraction. Training on Twitter dataset had a beneficial effect for DailyStrength model only in case of ADRs. This establishes a new state-of-the-art performance over the previous ADRMine system which achieved F1= 0.82 on DailyStrength and F1 = 0.72 on Twitter datasets for ADR extraction.

The normalization model was evaluated on the Tw-Resolve dataset used in the SMM4H 2019 shared task [22]. It achieved an end to end performance of F1-score 0.49 on the NER task and 0.35 on the end-to-end task beating the previous best systems at 0.46 and 0.34 to set a new state-of-the-art on the end-to-end entity extraction and normalization tasks. Based on submissions in the shared task we believe that incorporating other corpora might further benefit the extraction performance on both the NER and the normalization task.

The tool presented in this work, unlike prior work, is generic enough such that it an be used in isolation on social media posts regardless of whether the post also includes a symptom mention or drug mention. However, the performance of the tool has been evaluated on datasets that have higher incidence of symptoms in a collection of posts compared to collections from Twitter. Hence, for such datasets where the occurrence of symptoms are low, additional noise removal and filtering strategies such as the use of regular expressions, rules and/or supervised classifiers are recommended.

4 Conclusion

In this work we present a system to extract and normalize disease symptoms on social media posts. On a filtered balanced dataset, it obtained state-of-the-art performance for extracting symptoms, outscoring the ADRMine system by 5 percentage points on the Twitter dataset and 7 percentage points on the DailyStrength dataset for the ADE category. We have made SEED publicly available to users as a standalone tool, and include a web-based demonstration interface and an application programming interface which performs symptom extraction and normalization tasks on user submitted content as shown in 3.

Acknowledgements

A portion of this work was done when AM was a graduate student at Arizona State University supervised by GG and MS. MedDRA® the Medical Dictionary for Regulatory Activities terminology is the international medical terminology developed under the auspices of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). MedDRA® trademark is registered by IFPMA on behalf of ICH.

Funding

The work at University of Pennsylvania was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) grant R01LM011176 awarded to GG.

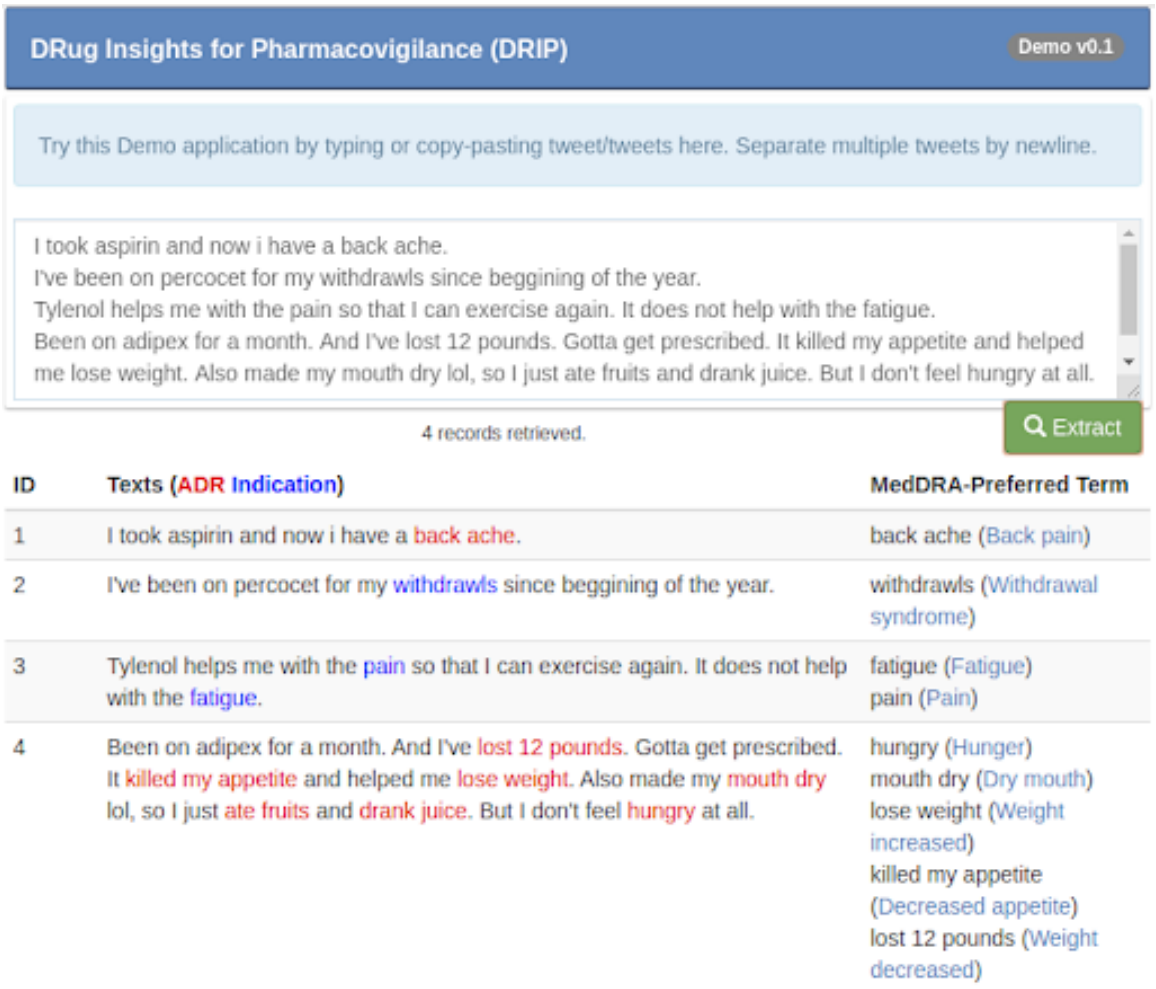

Figure 3: Screenshot of the DRIP System Demonstrating the Extraction of Symptoms from Social Media Texts.

References

[1] Xu Du, Onyeka Emebo, Aparna Varde, Niket Tandon, Sreyasi Nag Chowdhury, and Gerhard Weikum. Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning. In *2016 IEEE 32nd International Conference on Data Engineering Workshops (ICDEW)*, pages 54–59. IEEE, 2016.

[2] Siqi Zheng, Jianghao Wang, Cong Sun, Xiaonan Zhang, and Matthew E Kahn. Air pollution lowers chinese urbanites' expressed happiness on social media. *Nature Human Behaviour*, 3(3):237–243, 2019.

[3] Abeed Sarker, Annika DeRoos, and Jeanmarie Perrone. Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework. *Journal of the American Medical Informatics Association*, 27(2):315–329, 2020.

[4] Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015.

[5] Su Golder, Karen O'Connor, Sean Hennessy, Robert Gross, and Graciela Gonzalez-Hernandez. Assessment of beliefs and attitudes about statins posted on twitter: A qualitative study. *JAMA Network Open*, 3(6):e208953–e208953, 2020.

[6] Su Golder, Karen Smith, Karen O'Connor, Robert Gross, Sean Hennessy, and Graciela Gonzalez-Hernandez. A comparative view of reported adverse effects of statins in social media, regulatory data, drug information databases and systematic reviews. *Drug Safety*, pages 1–13, 2020.

[7] Karen Smith, Su Golder, Abeed Sarker, Yoon Loke, Karen O'Connor, and Graciela Gonzalez-Hernandez. Methods to compare adverse events in twitter to faers, drug information databases, and systematic reviews: proof of concept with adalimumab. *Drug safety*, 41(12):1397–1410, 2018.

[8] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth international AAAI conference on weblogs and social media*. Citeseer, 2011.

[9] Ashlynn R Daughton, Michael J Paul, and Rumi Chunara. What do people tweet when they're sick? a preliminary comparison of symptom reports and twitter timelines. In *ICWSM Social Media and Health Workshop*, 2018.

[10] Paola Velardi, Giovanni Stilo, Alberto E Tozzi, and Francesco Gesualdo. Twitter mining for fine-grained syndromic surveillance. *Artificial intelligence in medicine*, 61(3):153–163, 2014.

[11] Jia-Wen Guo, Christina L Radloff, Sarah E Wawrzynski, and Kristin G Cloyes. Mining twitter to explore the emergence of covid-19 symptoms. *Public Health Nursing*, 2020.

[12] Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. Self-reported covid-19 symptoms on twitter: An analysis and a research resource. *medRxiv*, 2020.

[13] Juan M Banda, Gurdas Viguruji Singh, Osaid Alser, and DANIEL PRIETO-ALHAMBRA. Long-term patient-reported symptoms of covid-19: an analysis of social media data. *medRxiv*, 2020.

[14] Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J. Am. Med. Inform. Assoc.*, 22(3):671–681, May 2015.

[15] Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J Paul, and Graciela Gonzalez-Hernandez. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019, 2019.

[16] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[17] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[18] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, December 2017.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. October 2018.

[20] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

[21] O Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology, 2004.

[22] Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez-Hernandez. 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*, 2019.