A Literature Review on Medication Names Extraction from Social Media and ADR Classification

Mia Mohammad Imran
Virginia Commonwealth University
Richmond, Virginia, U.S.A.
imranm3@vcu.edu

Abstract

Social media posts have become an enormous source of population health. Users often share their experiences about medication intake and their effects. A fundamental step toward incorporating social media data in pharmacoepidemiologic research is to automatic medication names detection in these posts. However, detecting drug name mentioning posts are challenging as they are often informal and noisy. Medication names are often misspelled and abbreviated. Besides, they are often context-dependent. The purpose of this study is to explore some of the aspects of medication name extraction and adverse drug reaction(ADR) from social media. However, the focus is mostly on tweets.

I. Introduction

Social media posts are often useful to detect real-time adverse drug reactions (ADR) compared to other available resources such as clinical reports, health records. They can be a valuable tool for pharmaceutical companies to obtain feedback about medication [1], [2]. Data can easily be collected in real-time. Plachouras et al. [1] reported that from July 9 to September 4, 2014 (shown in Figure 1), on average 179K tweets were mentioning drug names, and 721 tweets were containing ADRs. They proposed an ecosystem on clinical reports and social media posts can be useful to pharmaceutical companies - shown in Figure 2.

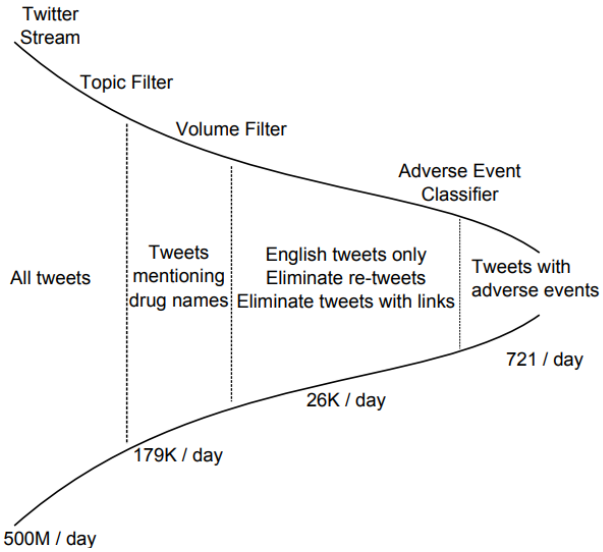

Fig. 1: Avg. daily tweets containing ADR from July 9 to Sep 4, 2014 by Plachouras et al. [1]

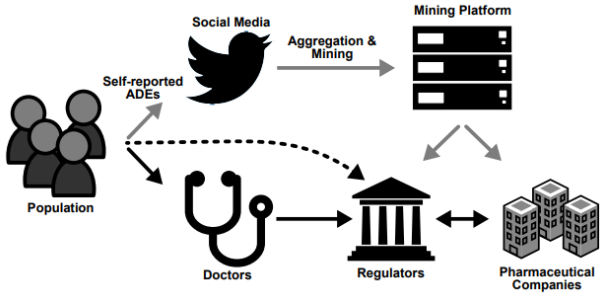

Fig. 2: The ecosystem of pharmaceutical companies, regulators, doctors, patients and social media Plachouras et al. [1].

The problem of detecting medication names and ADRs mentioning social media are often interrelated. Weissenbacher et al. [3] noted that these researches have been conducted extensively in recent years. Denecke et al. [4] performed a content analysis of medical social media data, in particular question & answer portals, weblogs, reviews, and wikis. De Rosa et al. [5] researched cross-relating social media content with trusted sources such as PubMed. The researchers have essentially framed the task of medication name extraction and ADR detection as a binary classification problem. Machine learning techniques are widely used in this task. Deep learning techniques are becoming popular of late. In this review, we will give short overview of various techniques used by researchers.

This review is divided into a couple of sections: preprocessing, model, annotation, evaluation.

II. Preprocessing

Pimpalkhute et al. [6] noted that preprocessing medication names in social media data often faces the problem of phonetic spelling. Their proposed model found that 50.4 – 56.0 % of the user comments using only about 18% of the variants (of spelling). The method, as shown in Figure 3, is based on the intuitive notion that people will tend to spell drug names phonetically. They have used tools and libraries such as the LOGIOS Lexicon Tool [1], Metaphone library [2], and CMU library [3]. The drug names are available in DrugBank [4]. Limsopatham et al. [7] used a phrase-based machine translation technique [8] to translate phrases from *Twitter language to formal medical language*.

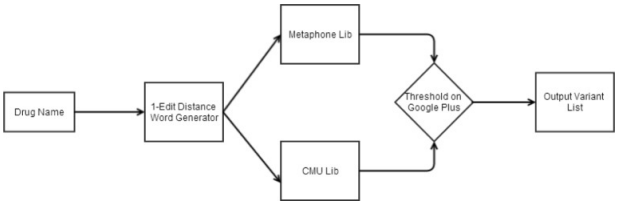

Fig. 3: Flow chart of Pimpalkhute et al. [6] methodology.

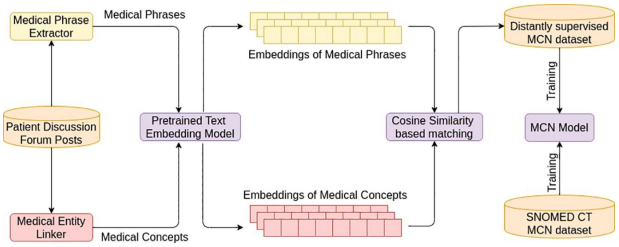

Fig. 4: Distant supervision for MCN model by Pattisapu et al. [9].

Limsopatham et al. [10] later proposed two different architecture for medical concept normalization (MCN) - a Convolutional Neural Network (CNN) model (Figure 6) and a Recurrent Neural Network (RNN) model (Figure 5). Pattisapu et al. [9] proposed MCN maps to translate from informal phrases to formal medical concepts. Their proposed model is highlighted in Figure 4.

III. Model

A. Machine Learning

One of the earliest research about medication names detection in social media was conducted by Jimeno-Yepes et al. [11]. They used two off-the-shelf classifiers - MetaMap [5], and the Stanford NER tagger [6], and a method based on conditional random fields. Jonnagaddala et al. [12], as a part of PSB 2016 Social Media Mining shared task [7], developed a methodology using SVM to classify whether tweets contains ADR or not. Sarker et al. [13] proposed a stacking based ensemble classifier to automatically

[1] http://www.speech.cs.cmu.edu/tools/lextool.html
[2] https://en.wikipedia.org/wiki/Metaphone
[3] https://en.wikipedia.org/wiki/CMUPronouncingDictionary
[4] https://go.drugbank.com/stats
[5] https://metamap.nlm.nih.gov/
[6] https://nlp.stanford.edu/software/CRF-NER.html
[7] http://diego.asu.edu/psb2016/task2data.html

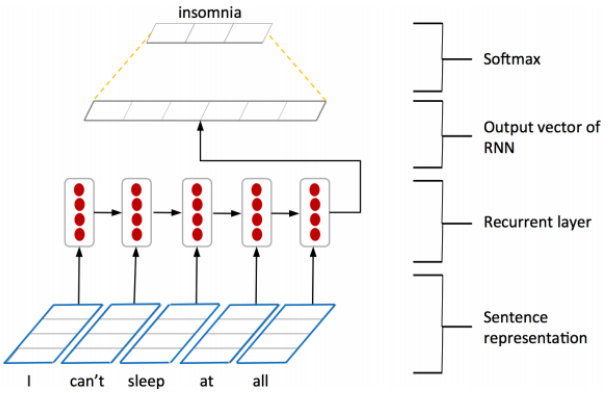

Fig. 5: RNN model for MCN model by Limsopatham et al. [10].

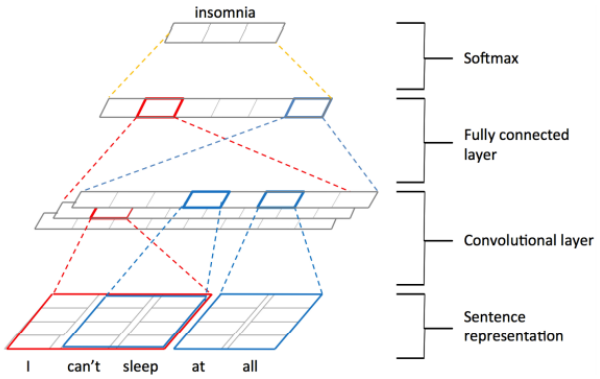

Fig. 6: CNN model for MCN model by Limsopatham et al. [10].

detect whether tweets are medication abuse or not for specific 4 medications. The architecture of the methodology is shown in Figure 7. In the ensemble classifier, they used four algorithms - naive bayes (NB), SVM, maximum entropy (ME), and J48, and used various features such as n-gram, abuse-indicating terms, lexicon matches, synonym expansion, word cloud. They concluded that tweets are often ambiguous and impersonal, and a large number of corpus needed for better results. A sample of their annotated data is available [8]. Meanwhile, Zhang et al. [14] applied an ensemble algorithm of four classifiers to detect a tweet contains ADR or not: (1) a concept-matching classifier based on ADR lexicon; (2) a ME classifier with word-level n-gram features and TFIDF weighting scheme; (3) a ME classifier based on word-level n-grams using NB log count ratios as feature values; and (4) a ME classifier with word embedding features. The code of their model is available [9]. Dai et al. [15] applied SVM to classify ADR posts and features such as linguistic, polarity, lexicon, and topic modeling based features. The resources applied for this research is available [10]. Alimova et al. [16] applied two types of feature engineering(context-level and entity-level) and applied two machine learning based models - Linear SVM and Logistic Regression (LR). The source code of this model is available [11].

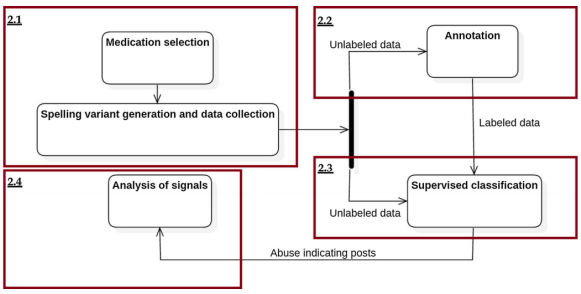

Fig. 7: Architecture by Sarker et al. [13]

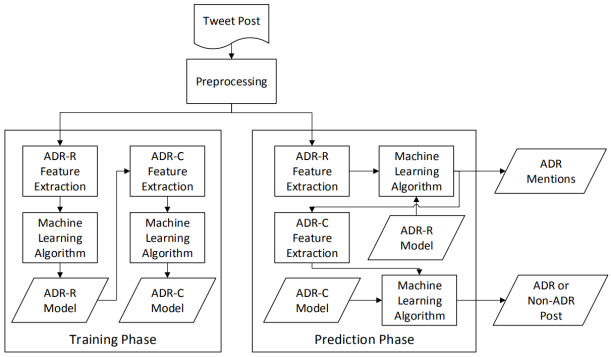

Fig. 8: Flow chart of Dai et al. [15]'s methodology.

[8]http://diego.asu.edu/Publications/DrugAbuse_DrugSafety.html
[9]https://github.com/tjflexic/psb-adr
[10]https://sites.google.com/site/hjdairesearch/Projects/adverse-drug-reaction-mining
[11]https://github.com/Ilseyar/adr_classification

These machine learning based approaches [12]–[16] heavily rely on feature engineering and require large amount of expert knowledge.

B. Deep Learning

Recently the focus has shifted towards deep learning. Tutubalina et al. [17] proposed a model based on bidirectional RNN for MCN in social media. The architecture is illustrated in Figure 9. Huynh et al. [2] investigated different neural network (NN) architecture for ADR classification. In particular, they applied their data over four algorithms: CNN, Recurrent Convolutional Neural Network (RCNN), Convolutional Recurrent Neural Network (CRNN), Convolutional Neural Network with Attention (CNNA). The architecture of these algorithms are shown in the Figure 7. The source code of the project is available [12]. Lee et al. [18], meanwhile, applied a semi-supervised CNN framework for the same task. This model was developed by Johnson et al. [19]. Lee proposed a Semi-supervised CNN model. The architecture of the model is shown in the Figure 10. The method works in two phases: (1) unsupervised phrase embedding learning, and (2) integrating the learned embeddings into the supervised training that uses labeled data. Wu et al. [20], [21] proposed neural approach using multi-head self-attention (MSA) to jointly detect drug name and adverse drug reaction. Their MSA model has three modules: first, a word representation module, which aims to build the contextual representations of words from the original characters within them; second, a tweet representation module; third, a classification module. The framework of the model is illustrated in the Figure 13.

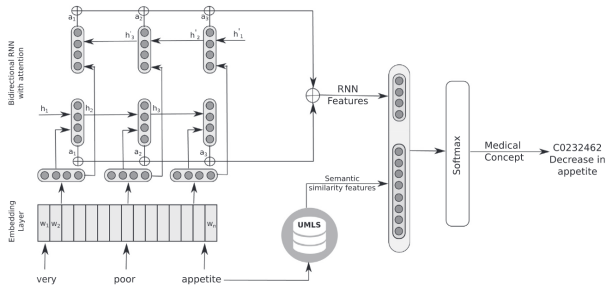

Fig. 9: Architecture for MCN in social media by Tutubalina et al. [17].

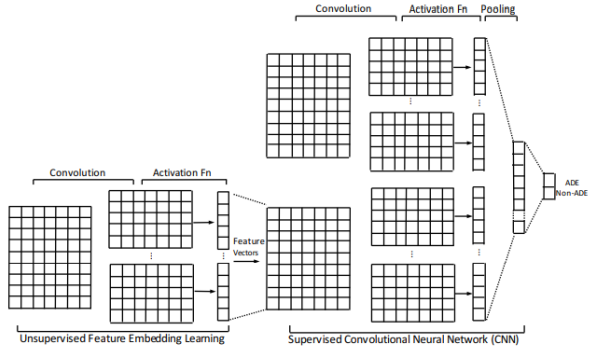

Fig. 10: Semi-supervised CNN by Lee et al. [18].

Weissenbacher et al. [22] introduced a deep neural networks ensemble method - *Kusuri* - a LSTM model - to detect medication names in unbalanced tweets. *Kusuri* is composed of two modules: first, applied four different classifiers (lexicon based, spelling variant based, pattern based, and a weakly trained neural network) parallel to discover potential tweets with medication names; second, an ensemble of deep neural networks encoding morphological, semantic, and longrange dependencies of important words in the tweets makes the final decision. The architecture is shown in Figure 12. A brief description of four algorithms are provided by the authors [13].

Both supervised and unsupervised models have been developed based on neural networks architecture. CNN [2], [18], [21] and RNN [2], [17] based models have been popular among researchers.

[12]https://github.com/trunghlt/AdverseDrugReaction

[13]https://tinyurl.com/y47qzbcq

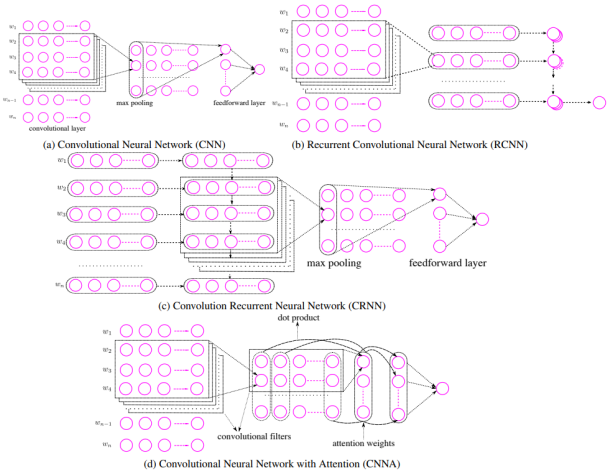

Fig. 11: Architecture by Huynh et al. [2].

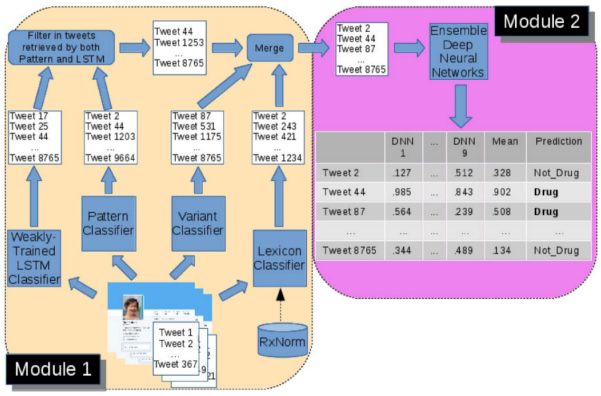

Fig. 12: *Kusuri* model by Weissenbacher et al. [22].

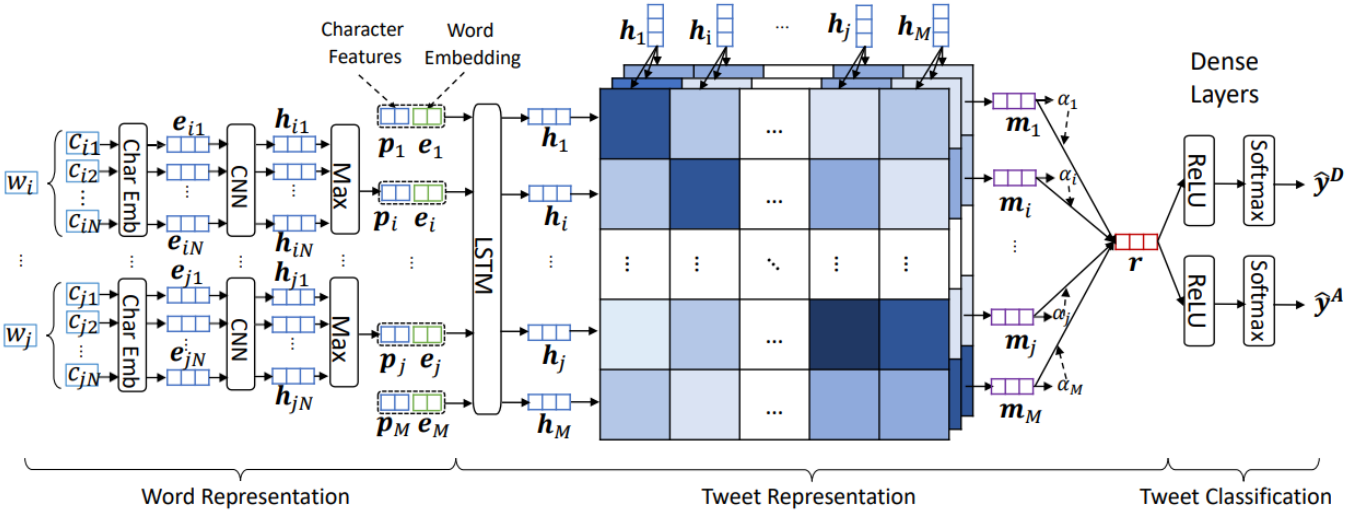

Fig. 13: MSA model by Wu et al. [20], [21].

IV. Annotation and Corpus

Annotating tweets have been one of the most difficult tasks. In their proposed model, Alvaro et al. [23] annotated over 1548 tweets in three categories: “First-class experience”, “Tweet written in English language”, and “Tweet about the drug”. There curated corpus is available [14]. The ADR expert guideline is available as well [15].

Sarker et al. [13] in their follow-up work, proposed a corpus of 267215 Twitter posts containing at least one drug related keyword [24]. They collected this data over a four-month period - from November, 2014 to February, 2015. This data contain over 250 medication related keywords. This data is available in their website [16]. Alvaro et al. [25] presented a corpus of pharmacovigilance which compromised of 1000 tweets and 1000 PubMed sentences. For the task, they collected 165489 tweets and 29435 PubMed sentences. Their TwiMed pipeline annotation is shown in the Figure 14.

[14]http://github.com/nestoralvaro/JBI_Pharmacovigilance
[15]https://ars.els-cdn.com/content/image/1-s2.0-S1532046415002415-mmc1.pdf
[16]http://diego.asu.edu/Publications/Drugchatter.html

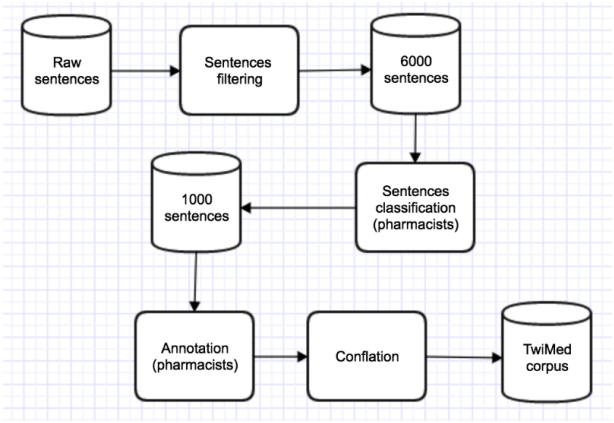

Fig. 14: TwiMed annotation pipeline by Alvaro et al. [25].

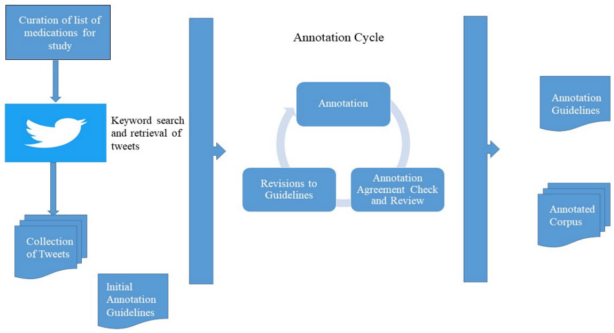

Fig. 15: Overview of the creation of the annotation guideline by O'Connor et al. [26].

Klein et al. [27] focused their study not only on detecting medications but also on distinguishing tweets whether they indicate the user possibly took it or merely mention medication. For this research, they annotated 10260 tweets. The annotation guidelines and a sample of the annotated data are available [17]. Later, Klein et al. [28] proposed a corpus of 27941 tweets for medication intake classification. They annotated them in three categories - "intake", "possible intake", and "no intake". O'Connor et al. [26] proposed an annotation guideline when they annotated 16443 tweets mentioning at least 20 abuse-prone medications. The guideline process is shown in Figure 15. They categorized tweets into four categories: "potential abuse or misuse", "non-abuse consumption", "drug mention only", and "unrelated". The annotation guideline is available [18].

V. Evaluation

Over the years, different techniques have been applied for preprocessing, model development, and annotation. Machine learning approaches have been widely used. Deep learning techniques are gradually replacing them and becoming more successful to detect medication names and ADR events.

For ADR classification, among the machine learning-based models, Jonnagaddala et al. [12], in their binary classification, managed to achieve an F-score of 0.33. Meanwhile, Zhang et al. [14]'s ensemble method obtained an F-score of 0.4182 and Sarker et al. [13]'s ensemble method obtained an F-score of 0.46. Huynh et al. [2] compared their deep learning-based model against Zhang et al. [14]'s model. On twitter dataset, their [2] model achieved an F-score of 0.51, 0.51, 0.49, and 0.49 using CNN, CRNN, RCNN, and CNNA respectively compared to baseline model's 0.40 F-score. On an unbalanced dataset, Lee et al. [18]'s Semi-Supervised CNN model for ADR classification outperforms the state-of-the-art supervised ADR classifier from previous work [29] by +9.9%. Wu et al. [21] compared their MSA model with Lee et al. [18]. On a twitter dataset, where Lee et al.' [18]s model achieved an F-score of 0.503, Wu et al. [21] managed to obtain 0.524. However, Alimova et al. [16] showed that it's possible to achieve better results using machine learching techniques compared to neural networks when the model is feature-rich. On a twitter dataset, their model [16] achieved an F-score of 0.702 using SVM and 0.737 using LR against a CNN baseline which achieved 0.702.

Meanwhile, on the research of detecting medication names, Jimeno-Yepes et al. [11] achieved an F-score of 0.65 on a corpus of 1300 tweets using a method based on conditional random fields. Wu et a. [20] achieved an F-score of 0.9183 using their MSA model on a balanced dataset. Weissenbacher

[17]https://healthlanguageprocessing.org/twitter-med-intake/
[18]https://tinyurl.com/y2zetam9

et al. [22] obtained an F-score of 0.788 on an extremely unbalanced dataset (98959 tweets, with only 0.26% mentioning medications) and 0.937 on a balanced dataset(50-50 corpus of 15005 tweets) using deep learning methodologies.

VI. Conclusion

The research of extracting medication names and detecting ADR has been progressed rapidly during the last couple of years. Machine learning and neural network has been the basis of these proposed models. However, the researchers noted that two major issues they have been facing: the problem of an unbalanced dataset and that the process of annotation.

References

[1] V. Plachouras, J. L. Leidner, and A. G. Garrow, "Quantifying self-reported adverse drug events on twitter: signal and topic analysis," in *Proceedings of the 7th 2016 International Conference on Social Media & Society*, 2016, pp. 1–10.

[2] T. Huynh, Y. He, A. Willis, and S. Rüger, "Adverse drug reaction classification with deep neural networks." Coling, 2016.

[3] D. Weissenbacher, A. Sarker, M. Paul, and G. Gonzalez, "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*, 2018, pp. 13–16.

[4] K. Denecke and W. Nejdl, "How valuable is medical social media data? content analysis of the medical web," *Information Sciences*, vol. 179, no. 12, pp. 1870 – 1880, 2009, special Section: Web Search. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025509000528

[5] M. De Rosa, G. Fenza, A. Gallo, M. Gallo, and V. Loia, "Pharmacovigilance in the era of social media: Discovering adverse drug events cross-relating twitter and pubmed," *Future Generation Computer Systems*, vol. 114, pp. 394–402.

[6] P. Pimpalkhute, A. Patki, A. Nikfarjam, and G. Gonzalez, "Phonetic spelling filter for keyword selection in drug mention mining from social media," *AMIA Summits on Translational Science Proceedings*, vol. 2014, p. 90, 2014.

[7] N. Limsopatham and N. Collier, "Adapting phrase-based machine translation to normalise medical terms in social media messages," *arXiv preprint arXiv:1508.02285*, 2015.

[8] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, Tech. Rep., 2003.

[9] N. Pattisapu, V. Anand, S. Patil, G. Palshikar, and V. Varma, "Distant supervision for medical concept normalization," *Journal of Biomedical Informatics*, vol. 109, p. 103522, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046420301507

[10] N. Limsopatham and N. Collier, "Normalising medical concepts in social media texts by learning semantic representation." Association for Computational Linguistics, 2016.

[11] A. Jimeno-Yepes, A. MacKinlay, B. Han, and Q. Chen, "Identifying diseases, drugs, and symptoms in twitter." 2015.

[12] J. Jonnagaddala, T. R. Jue, and H.-J. Dai, "Binary classification of twitter posts for adverse drug reactions," in *Proceedings of the social media mining shared task workshop at the pacific symposium on biocomputing, Big Island, HI, USA*. PSB, 2016, pp. 4–8.

[13] A. Sarker, K. O'Connor, R. Ginn, M. Scotch, K. Smith, D. Malone, and G. Gonzalez, "Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter," *Drug safety*, vol. 39, no. 3, pp. 231–240, 2016.

[14] Z. Zhang, J. Nie, and X. Zhang, "An ensemble method for binary classification of adverse drug reactions from social media," in *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016, p. 1.

[15] H.-J. Dai, M. Touray, J. Jonnagaddala, and S. Syed-Abdul, "Feature engineering for recognizing adverse drug reactions from twitter posts," *Information*, vol. 7, no. 2, p. 27, 2016.

[16] I. Alimova and E. Tutubalina, "Automated detection of adverse drug reactions from social media posts with machine learning," in *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 2017, pp. 3–15.

[17] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, and V. Malykh, "Medical concept normalization in social media posts with recurrent neural networks," *Journal of Biomedical Informatics*, vol. 84, pp. 93 – 102, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046418301126

[18] K. Lee, A. Qadir, S. A. Hasan, V. Datla, A. Prakash, J. Liu, and O. Farri, "Adverse drug event detection in tweets with semi-supervised convolutional neural networks," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 705–714.

[19] R. Johnson and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *Advances in neural information processing systems*, 2015, pp. 919–927.

[20] C. Wu, F. Wu, J. Liu, S. Wu, Y. Huang, and X. Xie, "Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention," in *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, 2018, pp. 34–37.

[21] C. Wu, F. Wu, Z. Yuan, J. Liu, Y. Huang, and X. Xie, "Msa: Jointly detecting drug name and adverse drug reaction mentioning tweets with multi-head self-attention," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 33–41.

[22] D. Weissenbacher, A. Sarker, A. Klein, K. O'Connor, A. Magge, and G. Gonzalez-Hernandez, "Deep neural networks ensemble for detecting medication mentions in tweets," *Journal of the American Medical Informatics Association*, vol. 26, no. 12, pp. 1618–1626, 2019.

[23] N. Alvaro, M. Conway, S. Doan, C. Lofi, J. Overington, and N. Collier, "Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use," *Journal of biomedical informatics*, vol. 58, pp. 280–287, 2015.
[24] A. Sarker and G. Gonzalez, "A corpus for mining drug-related knowledge from twitter chatter: Language models and their utilities," *Data in brief*, vol. 10, pp. 122–131, 2017.
[25] N. Alvaro, Y. Miyao, and N. Collier, "Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations," *JMIR public health and surveillance*, vol. 3, no. 2, p. e24, 2017.
[26] K. O'Connor, A. Sarker, J. Perrone, and G. G. Hernandez, "Promoting reproducible research for characterizing nonmedical use of medications through data annotation: Description of a twitter corpus and guidelines," *Journal of medical Internet research*, vol. 22, no. 2, p. e15861, 2020.
[27] A. Klein, A. Sarker, M. Rouhizadeh, K. O'Connor, and G. Gonzalez, "Detecting personal medication intake in twitter: an annotated corpus and baseline classification system," in *BioNLP 2017*, 2017, pp. 136–142.
[28] A. Z. Klein, A. Sarker, K. O'Connor, and G. Gonzalez-Hernandez, "An analysis of a twitter corpus for training a medication intake classifier," *AMIA Summits on Translational Science Proceedings*, vol. 2019, p. 102, 2019.
[29] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *Journal of biomedical informatics*, vol. 53, pp. 196–207, 2015.