MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220182

Extraction of Medication-Effect Relations in Twitter Data with Neural Embedding and Recurrent Neural Network

Keyuan Jiang[a], Dingkai Zhang[b], Gordon R. Bernard[c]

[a] *Department of Computer Information Technology & Graphics, Purdue University Northwest, Hammond, Indiana, U.S.A.*
[b] *School of Information and Intelligent Engineering, Ningbo City College of Vocational Technology, Ningbo, Zhejiang, China*
[c] *Department of Medicine, Vanderbilt University, Nashville, Tennessee, U.S.A.*

Abstract

Recently, an active area of research in pharmacovigilance is to use social media such as Twitter as an alternative data source to gather patient-generated information pertaining to medication use. Most of thr published work focuses on identifying mentions of adverse effects in social media data but rarely investigating the relationship between a mentioned medication and any mentioned effect expressions. In this study, we treated this relation extraction task as a classification problem, and represented the Twitter text with neural embedding which was fed to a recurrent neural network classifier. The classification performance of our method was investigated in comparison with 4 baseline word embedding methods on a corpus of 9516 annotated tweets.

Keywords:

Drug-Related Side Effects and Adverse Reactions, Deep Learning, Social Media

Introduction

Social media have become a popular platform for people to share their personal experiences, including those pertaining to their intake of medications. The wealth and value of personal medication experiences on social media have driven active research endeavors in using social media as an alternative data source for pharmacovigilance in order to augment the surveillance data. In 2015, Golder and colleagues [1] collected over 3000 publications germane to pharmacovigilance and social media. However, the primary focus has been on identifying mentions of pharmaceutical products and adverse drug reactions or events, which is a task of entity recognition in information retrieval. Very little has been done in understanding the relations between the mentions of drugs and their effects, and discovery of such relations remains largely a manual process, as reported in the work of identifying potentially unreported effects of Humira and opioids from Twitter data [2,3].

There exist a number of methods for extracting drug-related relations in biomedical field, including machine learning-based methods [4-6], dependency tree-based approaches [7-10], and kernel-based methods [11-13]. These methods were developed mainly for processing formal writings of scientific literature.

Thanks to their abundance and relatively short form, Twitter data were considered for this study. As the posts on a general social media platform, Twitter data possess unique characteristics unfound in formal writings: they are noisy, they may contain creative short texts to include the needed information within the space limit, their textual content may not follow grammatical and spelling rules, and they can be ambiguous due to short text. All these make the conventional methods for formal writing perform poorly [22]. In addition, there is a lack of relation extraction tools for Twitter data, although there are published works on information extraction from Twitter data [14-16] which focused on entity recognition.

In identifying any relations in Twitter posts pertaining to medication effect, we defined three types of relation between a medication and an effect expression in a single tweet: side effect (s), indication (i) (for beneficial relation), and no relation (neither side effect nor indication) (n). This treatment of the medication effect relations follow the SIDER database [20] where only two types of relations, side effect and indication, were extracted from marketed drugs. The no-relation type was added to represent the cases where neither relation exists between a drug and an effect expression. Examples of these relations are as follows (medications are in boldface and effects are underscored):

> "I am the same. Been on **Tecfidera** 18 mo. Flushing and hives at first. Went away." (*s*)
>
> "**Abilify** - antipsychotic used to treat schizophrenia, bipolar disorder, and agitation." (*i*)
>
> "*shakes bottle of **ritalin***" (*n*)

The first tweet describes the side effects (flushing and hives) of Tecfidera, even though the medication and effects are in the same sentence. The second tweet clearly indicates that Abilify treats several disorders/effects. In the last tweet which is in our corpus of study tweets, word shack, which has multiple meanings (senses), may be interpreted as a synonym of the concept tremor. (http://linkedlifedata.com/resource/umls/id/C0040822). And word shack and drug Ritalin in this tweet has neither side effect nor indication relationship.

In this study, we treated the relation extraction as a supervised 3-class classification problem and investigated the classification performance of representing Twitter data using the neural embedding technique implemented in word2vec [17], which achieved the state-of-art results in many NLP tasks [17, 18]. In neural embedding, the language model is learned from unlabeled text data, and each term is represented as a dense vector of real numbers. Each tweet is represented as a serious of dense vectors which serve as the input of the subsequent classifier. The neural embedding language model embeds linguistic syntactic and semantic characteristics which can be leveraged for discovering relations between entities.

Methods

The pipeline of data processing and analysis is shown in Figure 1. After initial processing of Twitter data, which is described in the Data section, a corpus of tweets containing at least a medication and one or more effect expressions was generated. This corpus was annotated and later used by both baseline methods and the proposed neural embedding approach. Classification performance of each method was collected and evaluated. Finally, statistical analysis was conducted to confirm the existence of any differences in classification performance between the baseline methods and the proposed method.

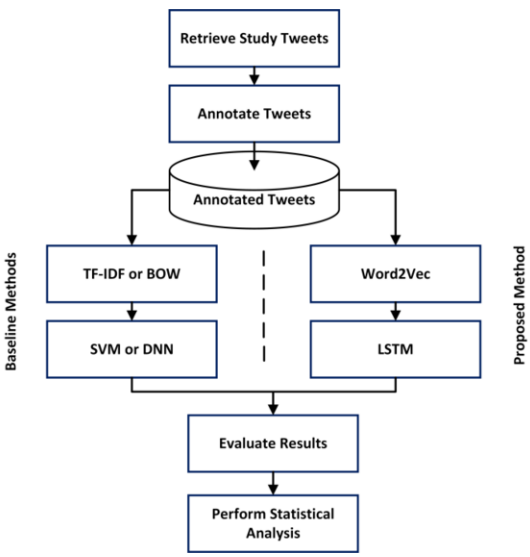

Figure 1 – Pipeline of Data Processing and Analysis

Data

A collection of 53 million tweets related to 100 medications was gathered using a home-made crawler in compliance with the Twitter.com access policy. These tweets were posted between March of 2006 and June of 2017, and collected in June of 2017. Preprocessing the 53 million raw tweets generated 12 million "clean" tweets which are English only, without duplicates and re-tweets. This corpus of "clean" tweets was used to train the Word2vec neural embedding language model, which represents each token in tweets as a dense vector of real numbers. Later, the "clean" tweets were filtered with effect terms obtained from the SIDER database [20], a side effect resource hosted at the European Molecular Biology Laboratory, and their variations (synonyms) in the consumer health vocabulary (CHV) [21], leading to a corpus of 3.6 million tweets which were used to infer the pairs of medication and effect based upon relational similarity. Finally, 300 pairs were randomly selected and 9516 tweets containing each pair of medication and effect were retrieved from the collection of 3.6 million tweets [2].

The 9516 selected tweets were annotated by two annotators according to an annotation guideline. Each tweet was given a label *s*, *i*, or *n*, representing side effect, indication, or no relation. Any disagreement in the annotation was resolved by another researcher. Table 1 lists the composition of the annotated tweet corpus.

Table 1 – Composition of the Annotated Tweet Corpus

Class	# of Tweets
Side Effect (*s*)	2470
Indication (*i*)	2299
No Relation (*n*)	4747
Total	9516

Baseline Methods

To evaluate the classification performance of the proposed neural embedding approach, four word embedding methods were considered as the baseline. This arrangement helped us compare the neural embedding method with other word embedding approaches. They are TF-IDF (term frequency–inverse document frequency) with SVM (support vector machine) classifier, BOW (bag of words) with SVM classifier, TF-IDF with DNN (deep neural network) classifier, and BOW with DNN classifier. The DNN classifier was configured as having 5 hidden layers with 1500 nodes in each layer.

Neural Embedding Approach

For our proposed neural embedding approach, a word2vec model was generated using a corpus of 12 million unlabeled tweets. In testing, each annotated tweet was fed to word2vec [17] to generate a series of dense vectors. Word2vec was configured to use its skip-gram architecture with a window size of 10 – that is, 10 words before and after the center word, to have sufficient context span embedding the semantical relationship of a medication mention and effect expression. Other notable parameters are the minimum count of 5 and dimension of

Table 2 – Classification Performance Results. The highest value of each measure is in boldface.

Method	Acc	Prec (*s*)	Prec (*i*)	Prec (*n*)	Rec (*s*)	Rec (*i*)	Rec (*n*)	F1 (*w*)
BOW+SVM	0.681	0.693	0.567	0.716	0.707	0.445	0.779	0.673
TF-IDF+SVM	0.679	**0.706**	0.570	0.700	0.693	0.399	**0.802**	0.667
BOW+DNN	0.682	0.677	0.553	0.738	**0.718**	0.511	0.742	0.678
TF-IDF+DNN	0.671	0.688	0.535	0.712	0.696	0.458	0.755	0.664
w2v+LSTM	**0.707**	0.691	**0.590**	**0.766**	0.714	**0.565**	0.767	**0.703**

Table 3 – Wilcoxon Signed Rank Test Results (α = 0.05 and n = 10).

	Acc	**Prec (*s*)**	**Prec (*i*)**	**Prec (*n*)**	**Rec (*s*)**	**Rec (*i*)**	**Rec (*n*)**	**F1 (*w*)**
BOW+SVM	Sig	Not Sig	Sig	Sig	Not Sig	Sig	Sig	Sig
TF-IDF+SVM	Sig	Not Sig	Not Sig	Sig	Not Sig	Sig	Sig	Sig
BOW+DNN	Sig	Not Sig	Sig	Sig	Not Sig	Sig	Sig	Sig
TF-IDF+DNN	Sig	Not Sig	Sig	Sig	Not Sig	Sig	Not Sig	Sig

300. The output of word2vec model was fed to a long short-term memory (LSTM) neural network, a recurrent neural network capable of persisting information previously processed, helping retain semantics within the context of each tweet. The LSTM classifier was configured to have an input layer of 128 units and one dense output layer with 3 units (for 3 classes).

Implementation

Scikit-learn Python library (https://scikit-learn.org) was used to implement SVM, and DNN, and Keras library (https://keras.io), which runs on top of TensorFlow (https://www.tensorflow.org), was utilized to implement LSTM.

Statistical Analysis

All the methods in this study were evaluated with 10-fold cross-validation. The annotated tweet corpus was partitioned into the same 10 folds for all the methods. The average value was calculated for each performance measure.

To confirm the existence of any differences of classification performance between the neural embedding method and each baseline method, Wilcoxon Signed Rank Test performed on each pair of a baseline method and the proposed approach for the same data partition. The beauty of the Wilcoxon Signed-Rank Test is that it does not have any assumption of the distribution of the data. The performance difference is considered in existence if the signed-rank value is not more than the critical value of the given α (0.05) and n (10) for this study.

Results

Table 2 lists the performance measures of each method for each class (type) of relations. The first 4 methods are the baseline methods and the last one (w2v+LSTM) is the neural embedding approach. In the table, Acc: accuracy, Prec: precision, Rec: recall, s: s class, i: i class, n: n class, and w: weighted. F1 is the geometric mean of precision and recall of a given class. The weighted F1 was used to take into the consideration the class-imbalance of the dataset – that is, three classes do not have the same number of tweets in our dataset.

Listed in Table 3 are the results of our statistical analysis of the Wilcoxon Signed Rank Test on each performance measure between each pair of a baseline method and the neural embedding method. In the table, Sig: significant, and Not Sig: not significant. Significant means that performance difference between each pair of methods does exist, and not significant indicates that the difference is due to chance. Underscored results correspond to those with the highest values in Table 2.

Discussion

As can be seen in Table 2, the neural embedding method achieved highest values in 5 out of 8 performance measures, whereas 2 baseline methods achieved highest values in the remaining 3 measures: precision and recall on *s* class, and recall on *n* class.

In practical applications, we are more interested in whether a method can predict side effect(s) relations and/or beneficial effects (i) relations correctly, and the performance on n class should not be a major concern.

For accuracy, a measure for all the classes together, our method achieved the best performance with the support of statistical confidence.

As to precision, a measure of the percentage of true (actual) positives in the predicted result for a given class – e.g., the fraction of actual s class tweets are in the predicted s class tweets, our approach achieved mixed results. For the s class, there is no statistic support to the differences between our proposed method and any baseline methods. In other words, our proposed method may perform equally well as any of the baseline methods. For the i class, the proposed method performed better (than BOW+SVM, BOW+DNN and TF-IDF+DNN) or identically well (to TF-IDF+SVM).

In regards to recall is a measure of the fraction of true positive tweets in our dataset included in the prediction for a given class – for example, the fraction of the number of the actual s class tweets in the predicted s class vs. the total number of s class tweets in our corpus of annotated tweets. Our proposed method outperformed all the baseline methods on the i class, but the statistical analysis does not confirm that the difference on the s class between our method and each of the baseline methods. In other words, the recalls on the s class by different methods showed no difference.

The results of the weighted F1 score indicated that our method is superior to all the baseline methods with the statistic support, supporting that our method had the best performance.

In short, if a single method need to be considered, the neural embedding method appears to be the best candidate for our task.

To understand why and how our neural embedding method misclassified tweets for both s and i types (classes) of relations and to help improve the algorithm, we randomly sampled 40 misclassified tweets. Twenty of them were supposed to belong to the s class and the other 20 to the i class. For the 20 s class tweets, 10 were predicted for the i class and other 10 for n class, represent the misclassified s class tweets. The 20 i class tweets contained 10 predicted s class tweets and 10 predicted n class tweets.

There are a number of observations of misclassification. First, several tweets contain multiple medications and/or multiple effects. For example, "#Januvia and #Janumet have an associated risk of acute #pancreatitis when used to treat #diabetes" is an s class tweet for the Januvia-acute_pancreatitis. Not only does the tweet mention multiple drugs and effects, but it also contains both side effect (acute pancreatitis) and indication (diabetes) relations.

Some tweets seem to be mis-labeled. For instance, "@PERSON need a script but Nasonex works really well. Use Telfast as your antihistamine, non drowsy. Hope you feel better soon" was incorrected labeled as i class for the Nasonex-sleepiness relation. Here "non drowsy" is related to Telfast, not Nasonex. In addition, "non drowsy" is not sleepiness. This type of error is more

likely to be caused by our annotation process where the resolver only resolved disagreed labels without reviewing any tweets with agreed annotation by two annotators.

Twitter users tend to have a key hashtag at the end of their posts. In our case, the hashtag can be a medication or an effect. An example of such tweets is "My pills to treat my nausea ha side effects of nausea and shortness of breath which i have now haha at least im not bald yet! #omeprazole." This type of tweets can pose some difficulty of resolving coreference by computer.

Acronyms are common in medical terminology, and many can have multiple meanings. This situation may confuse a classifier. In this instance of the i class, "@PERSON (((hugs))) I hope its ok - PEs are a shit but very treatable, though clexane. :(:(:(" for the clexane-lung_embolism relation, PE means pulmonary embolism in the context, and our classifier predicted it as no relation (n).

Several side effect tweets (s class) were misclassified as indication (i class) tweets without obvious reasons. An example of such tweets is "just called my doctors office to tell them im having hives from adderall and that i also want to keep trying antidepressants" for the Adderall-hives relation.

For pharmacovigilance, tweets of personal experience related medication use are of special interest. Our ultimate goal is to automate discovery of self-reported experience of medication effects from abundant and ever-growing number of Twitter posts, to identify any potentially unreported medication effects. As such, a pipeline can be devised to first identify personal experience tweets [19] and afterwards extract medication-effect relations from the tweets. The outcome of this pipeline will be the tweets of personal experience related to medication use.

Conclusions

We studied the word2vec-based neural embedding of tweet text along with an LSTM recurrent neural network to extract medication-effect relations from medication-related tweets, and its performance was compared with 4 baseline methods. The results of the neural embedding method on a corpus of 9516 annotated tweets indicated that it outperformed the baseline methods in the majority of performance measures, and demonstrated no differences in other measures from our statistical analysis. This exhibits the utility of our neural embedding method for extracting medication-effect relations from Twitter posts. In short, the neural embedding method can be a good choice if a single method is needed for our task.

Acknowledgements

The authors wish to thank anonymous reviewers for their critiques and constructive comments that improved this manuscript, and Ravish Gupta, Liyuan Huang and Tingyu Chen for their work of identifying the pairs of medication and effect, as well as Gelareh Karbaschi and Youzhe Song for their efforts in annotation.

References

[1] S. Golder, G. Norman and Y. Loke, Systematic review on the prevalence, frequency and comparative value of adverse events data in social media, British *Journal of Clinical Pharmacology*, vol. 80, no. 4, 2015, pp. 878-888.

[2] K. Jiang, L. Huang, T. Chen, G. Karbaschi, D. Zhang and G. R. Bernard, Mining Potentially Unreported Effects from Twitter Posts through Relational Similarity: A Case for Opioids, *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Seoul, Korea (South), 2020, pp. 2603-2609.

[3] K. Jiang, S. Feng, L. Huang, T. Chen and G. R. Bernard, Mining Potential Effects of HUMIRA in Twitter Posts Through Relational Similarity, *Studies in health technology and informatics*, vol. 270, 2020, pp. 874-878.

[4] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki and K. Ohe, Extraction of adverse drug effects from clinical records, *MedInfo*, vol. 160, 2010, pp. 739-743.

[5] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius and L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, *Journal of biomedical informatics*, vol. 45, no. 5, 2012, pp. 885-892.

[6] Y. Zhang and Z. Lu, Exploring semi-supervised variational autoencoders for biomedical relation extraction, *Methods*, vol. 166, 2019, pp. 112-119.

[7] T. C. Rindflesch and M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, *Journal of biomedical informatics*, vol. 36, no. 6, 2003, pp. 462-477.

[8] A. Culotta and J. Sorensen, Dependency Tree Kernels for Relation Extraction, in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, 2004, pp. 423–429.

[9] C. Wang and F. James, Medical relation extraction with manifold models, In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 2014, pp. 828-838.

[10] M. Song, K. C. Won, L. Dahee, E. H. Go and Y. K. Keun, PKDE4J: Entity and relation extraction for public knowledge discovery, *Journal of biomedical informatics*, vol. 57, 2015, pp. 320-332.

[11] I. Segura-Bedmar, P. Martínez and C. de Pablo-Sánchez, Using a shallow linguistic kernel for drug–drug interaction extraction, *Journal of Biomedical Informatics*, vol. 44, no. 5, 2011, pp. 789-804.

[12] S. Kim, H. Liu, L. Yeganova and W. Wilbur, Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach, *Journal of Biomedical Informatics*, vol. 55, pp. 23-30, 2015.

[13] C. Giuliano, A. Lavelli and L. Roman, Exploiting shallow linguistic information for relation extraction from biomedical literature, In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006.

[14] B. Kalina, L. Derczynski, A. Funk, M. Greenwood, D. Maynard and N. Aswani, Twitie: An open-source information extraction pipeline for microblog text, in In *Proceedings of the international conference recent advances in natural language processing (RANLP) 2013*, Hissar, Bulgaria, 2013, pp. 83-90.

[15] M. Hasby and M. L. Khodra, Optimal path finding based on traffic information extraction from Twitter, *International Conference on ICT for Smart Society*, Jakarta, Indonesia, 2013, pp. 1-5.

[16] D. Anggareska and A. Purwarianti, Information extraction of public complaints on Twitter text for bandung government, *2014 International Conference on Data*

and Software Engineering (ICODSE), Bandung, Indonesia, 2014, pp. 1-6.

[17] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, In *Proceedings of Workshop at ICLR*, Scottsdale, Arizona, 2013.

[18] T. Mikolov, W. T. Yih, and G. Zweig, Linguistic regularities in continuous space word representations, In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 2013, pp. 746-751.

[19] K. Jiang, S. Feng, Q. Song, R.A. Calix, R, M. Gupta, and G.R. Bernard, Identifying tweets of personal health experience through word embedding and LSTM neural network. *BMC bioinformatics*, 19(8), 67-74.

[20] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, The SIDER database of drugs and side effects, *Nucleic acids research*, vol. 44, no. D1, 2015, pp.D1075-D1079.

[21] Q. T. Zeng and T. Tse, Exploring and developing consumer health vocabularies, *Journal of the American Medical Informatics Association*, vol. 13, no. 1, 2006, pp.24-29.

[22] K. Jiang, D. Zhang, and G. R. Bernard, Mining Medication-Effect Relations from Twitter Data Using Pretrained Transformer Language Model, In *Proceedings of the 2021 Machine Learning for Pharma and Healthcare Applications (PharML 2021),* Bilbao, Basque Country, Spain, 2021.

Address for correspondence

Keyuan Jiang, Purdue University Northwest, 2200 169th Street, Hammond, IN 46323, U.S.A, kjiang@pnw.edu.