Figurative Usage Detection of Symptom Words to Improve Personal Health Mention Detection

Adith Iyer♠♣, Aditya Joshi♠, Sarvnaz Karimi♠, Ross Sparks♠, Cécile Paris♠

♠CSIRO Data61, Sydney, Australia
♣University of Queensland, Brisbane, Australia
adith.iyer@uq.net.au, {firstname.lastname}@csiro.au

Abstract

Personal health mention detection deals with predicting whether or not a given sentence is a report of a health condition. Past work mentions errors in this prediction when symptom words, *i.e.*, names of symptoms of interest, are used in a figurative sense. Therefore, we combine a state-of-the-art figurative usage detection with CNN-based personal health mention detection. To do so, we present two methods: a pipeline-based approach and a feature augmentation-based approach. The introduction of figurative usage detection results in an average improvement of 2.21% F-score of personal health mention detection, in the case of the feature augmentation-based approach. This paper demonstrates the promise of using figurative usage detection to improve personal health mention detection.

1 Introduction

The World Health Organisation places importance on gathering intelligence about epidemics to be able to effectively respond to them (World Health Organisation, 2019). Natural language processing (NLP) techniques have been applied to social media datasets for epidemic intelligence (Charles-Smith et al., 2015). An important classification task in this area is *personal health mention detection*: to detect whether or not a text contains a personal health mention (PHM). A PHM is a report that either the author or someone they know is experiencing a health condition or a symptom (Lamb et al., 2013). For example, the sentence '*I have been coughing since morning*' is a PHM, while '*Having a cough for three weeks or more could be a sign of cancer*' is not. The former reports that the author has a cough while, in the latter, the author provides information about coughs in general. Past work in PHM detection uses classification-based approaches with human-engineered features (Lamb

et al., 2013; Yin et al., 2015) or word embedding-based features (Karisani and Agichtein, 2018). However, consider the quote '*When Paris sneezes, Europe catches cold*' attributed to Klemens von Metternich[1]. The quote contains names of symptoms (referred to as '*symptom words*' hereafter) '*sneezes*' and '*cold*'. However, it is not a PHM, since the symptom words are used in a figurative sense. Since several epidemic intelligence tools based on social media rely on counts of keyword occurrences (Charles-Smith et al., 2015), figurative sentences may introduce errors. Figurative usage has been quoted as a source of error in past work (Jimeno Yepes et al., 2015; Karisani and Agichtein, 2018). In this paper, we deal with the question:

> *Does personal health mention detection benefit from knowing if symptom words in a text were used in a literal or figurative sense?*

To address the question, we use a state-of-the-art approach that detects idiomatic usage of words (Liu and Hwa, 2018). Given a word and a sentence, the approach identifies if the word is used in a figurative or literal sense in the sentence. We refer to this module as '*figurative usage detection*'. We experiment with alternative ways to combine figurative usage detection with PHM detection, and report results on a manually labeled dataset of tweets.

2 Motivation

As the first step, we ascertain if the volume of figurative usage of symptom words warrants such attention. Therefore, we randomly selected 200 tweets (with no duplicates and retweets) posted in November 2018, each containing either 'cough' or 'breath'. After discarding tweets with garbled text,

---

[1] https://bit.ly/2VoqTif ; Accessed on 23rd April, 2019.

two annotators manually annotated each tweet with the labels 'figurative' or 'literal' to answer the question: '*Has the symptom word been mentioned in a figurative or literal manner?*'. Note that, (a) in the tweet '*When it's raining cats and dogs and you're down with a cough!*', the symptom usage is literal, and (b) Hyperbole (for example, '*soon I'll cough my entire lungs up*') is considered to be literal. The two annotators agreed on a label 93.96% of the time. Cohen's kappa coefficient for inter-rater agreement is 0.8778, indicating a high agreement. For 52.75% of these tweets, both annotators assign the label as figurative. This provides only an estimate of the volume of figurative usage of symptom words. We also expect that the estimate would differ for different symptom words.

3 Approach

We now introduce the approaches for figurative usage and PHM detection. Following that, we present two alternative approaches to interface figurative usage detection with PHM detection: the pipeline approach and the feature augmentation approach.

3.1 Figurative Usage Detection

In the absence of a health-related dataset labeled with figurative usage of symptom words, we implement the unsupervised approach to detect idioms introduced in Liu and Hwa (2018). This forms the figurative usage detection module. The input to the figurative usage detection module is a target keyword and a sentence, and the output is whether or not the keyword is used in a figurative sense. The approach can be summarised in two steps: computation of a literal usage score for target keyword followed by a LDA-based estimator to predict the label. To compute the literal usage score, Liu and Hwa (2018) first generate a set of words that are related to the target keywords (symptom words, in our case). This set is called the 'literal usage representation'. The literal usage score is computed as the average similarity between the words in the sentence and the words in the literal usage representation. Thus, this score is a real value between 0 and 1 (where 1 is literal and 0 is figurative). The score is then concatenated with linguistic features (described later in this section). The second step is a Latent Dirichlet Allocation (LDA)-based estimator. The estimator computes two distributions: the word-
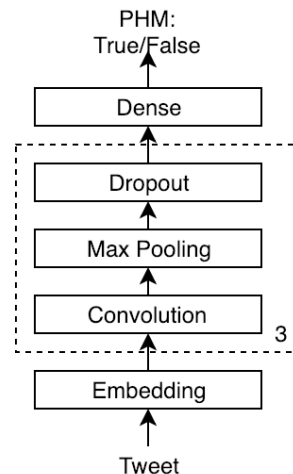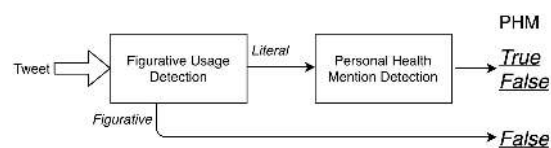

Figure 1: PHM detection.

Figure 2: Pipeline approach.

figurative/literal distribution which indicates the probability of a word to be either figurative or literal, and a document-figurative/literal distribution which gives a predictive score for a document to be literal or figurative. To obtain the literal usage score, we generate the literal usage representation using word2vec similarity learned from the Sentiment140 tweet dataset (Go et al., 2009). We use two sets of linguistic features, as reported in Liu and Hwa (2018): the presence of subordinate clauses and part-of-speech tags of neighbouring words, using Stanford CoreNLP (Manning et al., 2014). We adapt the abstractness feature in their paper to health-relatedness (*i.e.,* the presence of health-related words). The intuition is that tweets which contain more health-related words are more likely to be using the symptom words in a literal sense instead of figurative. Therefore, the abstractness feature in the original paper is converted to domain relatedness and captured using the presence of health-related words. We consider the symptom word as the target word. It must be noted that we do not have or use figurative labels in the dataset except for the sample used to report the efficacy of figurative usage detection.

3.2 PHM Detection

We use a CNN-based classifier for PHM detection, as shown in Figure 1. The tweet is converted

to its sentence representation using a concatenation of embeddings of the constituent words, padded to a maximum sequence length. The embeddings are initialised based on pre-trained word embeddings. We experiment with three alternatives of pre-trained word embeddings, as elaborated in Section 4. These are then passed to three sets of convolutional layers with max pooling and dropout layers. A dense layer is finally used to make the prediction.

3.3 Interfacing Figurative Usage Detection with PHM Detection

We consider two approaches to interface figurative usage detection with PHM detection:

1. **Pipeline Approach** places the two modules in a pipeline, as illustrated in Figure 2. If the figurative usage detection module predicts a usage as figurative, the PHM detection classifier is bypassed and the tweet is predicted to not be a PHM. If the figurative usage prediction is literal, then the prediction from the PHM detection module is returned. We refer to this approach as '+*Pipeline*'.

2. **Feature Augmentation Approach** augments PHM detection with figurative usage features. Therefore, the figurative label and the linguistic features from figurative usage detection are concatenated as figurative usage features ad passed through a convolution layer. The two are then concatenated in a dense layer to make the prediction. The approach is illustrated in Figure 3. This approach is based on Dasgupta et al. (2018), where they augment additional features to word embeddings of words in a document. We refer to this approach as '+*FeatAug*'.

In +Pipeline, the figurative label guides whether or not PHM detection will be called. In +FeatAug, the label becomes one of the features. For both the approaches, the figurative label is determined by producing the literal usage score and then applying an empirically determined threshold. We experimentally determine if using the literal usage score performs better than using the LDA-based estimator (See Section 4.3).
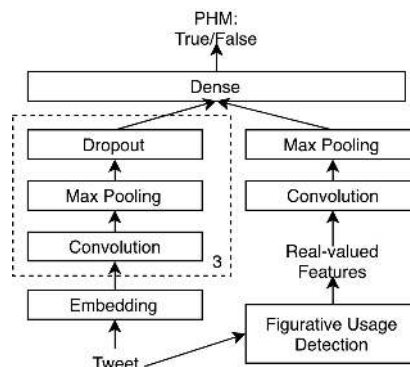

Figure 3: Feature augmentation approach.

4 Experiment Setup

4.1 Dataset

We report our results on a dataset introduced and referred to by Karisani and Agichtein (2018) as the PHM2017 dataset. This dataset consists of 5837 tweets related to a collection of diseases: Alzheimer's (1103, 16.7% PHM), heart attack (973, 12.4% PHM), Parkinson's (868, 9.8% PHM), cancer (988, 20.6% PHM), depression (924, 38.5% PHM) and stroke (981, 14.2% PHM). The imbalance in the class labels of the dataset must be noted. Some tweets in the original paper could not be downloaded due to deletion or privacy settings.

4.2 Configuration

For PHM detection (PHMD) and the two combined approaches (+Pipeline and +FeatAug), the parameters are empirically determined as:

1. **PHMD**: Filters=100, Kernels=(3, 4, 5), Pool size=2; Dropout=(0.2, 0.3, 0.5).

2. **Figurative Usage Detection**: The figurative label is predicted using a threshold for the literal usage score. This threshold is set to 0.2. This holds for both +Pipeline and +FeatAug. In the case of +Pipeline, a tweet is predicted as figurative, and, as a result, non-PHM, if the literal usage score is lower than 0.2. In the case of +FeatAug, the figurative label based on the score is added along with other features.

3. **+FeatAug**: Filters=100; Kernel size (left)=(3, 4, 5), Pool size=2; Dropout=(0.3, 0.1, 0.3); Kernel size (right)=2.

All experiments use the Adam optimiser and a batch size of 128, and trained for 35 epochs. CNN experiments use the ReLU activation. We

Approach	Random			word2vec			GloVe			Numberbatch		
	P	R	F	P	R	F	P	R	F	P	R	F
PHMD	59.40	39.84	46.31	57.85	47.24	50.99	68.71	50.70	57.05	59.07	43.59	48.63
+Pipeline	59.99	33.65	41.78	57.84	40.80	46.62	67.93	43.25	51.51	59.09	36.74	43.69
+FeatAug	54.51	45.01	**48.08**	57.11	51.71	**53.13**	66.70	53.52	**58.25**	54.48	48.75	**50.45**

Table 1: Performance of PHM Detection (PHMD), +Pipeline and +FeatAug for four word embedding initialisations. P: Precision, R: Recall, and F: F-score.

Approach	GloVe+MeSH			GloVe+WordNet			GloVe+Symptom		
	P	R	F	P	R	F	P	R	F
PHMD	56.95	41.47	46.62	56.41	42.94	47.55	57.57	42.93	47.72
+Pipeline	56.01	34.98	41.75	55.86	36.63	43.12	57.10	36.34	42.90
+FeatAug	53.71	46.46	**49.01**	55.88	48.47	**51.15**	56.04	48.11	**50.30**

Table 2: Performance of PHM Detection (PHMD), +Pipeline and +FeatAug initialised with GloVe word embeddings retrofitted with three ontologies: MeSH, WordNet and Symptom. P: Precision, R: Recall, and F: F-score.

	P	R	F	ΔF
PHMD	59.48	44.18	49.26	
+Pipeline	59.12	37.60	44.48	-4.78
+FeatAug	57.32	48.88	51.48	**+2.21**

Table 3: Average performance of PHM Detection (PHMD), +Pipeline and +FeatAug across the seven word embedding initialisations; P: Precision, R: Recall, F: F-score; ΔF: Difference in the F-score in comparison with PHMD.

Disease	PHMD	+FeatAug
Alzheimer's	65.33	**68.48**
Heart attack	**46.96**	45.98
Parkinson's	48.83	**51.49**
Cancer	53.69	**54.58**
Depression	70.48	**71.34**
Stroke	57.03	**57.65**

Table 4: Impact of figurative usage detection for PHM Detection (PHMD) on individual diseases.

use seven types of initialisations for the word embeddings. The first four are a random initialisation, and three pre-trained embeddings. The pre-trained embeddings are: (a) word2vec (Mikolov et al., 2013); (b) GloVe (trained on Common Crawl) (Pennington et al., 2014); and, (c) Numberbatch (Speer et al., 2017). The next three are embeddings retrofitted with three ontologies. We

use three ontologies to retrofit GloVe embeddings using the method by Faruqui et al. (2015). The ontologies are: (a) MeSH,[2] (b) Symptom[3], and (c) WordNet (Miller, 1995). The results are averaged across 10-fold cross-validation.

4.3 Evaluation of Figurative Usage Detection

To validate the performance of figurative usage detection, we use the dataset of tweets described in Section 2. The tweets contain symptom words that have been manually labeled. We obtain an F-score of (a) 76.46% when only the literal usage score is used, and (b) 69.72% when the LDA-based estimator is also used. Therefore, we use the literal usage score along with the figurative usage features for our experiments.

5 Results

The effectiveness of PHMD, +Pipeline and +FeatAug for the four kinds of word embedding initialisations is shown in Table 1. In each of these cases, +FeatAug performs better than PHMD, while +Pipeline results in a degradation. We note that, for both +FeatAug and +Pipeline, the recall is impacted in comparison with PHMD. Similar trends are observed for the retrofitted embeddings, as shown in Table 2. The improvement when figurative usage detection is used is

---

[2] https://www.nlm.nih.gov/mesh/meshhome.html; Accessed on 23rd April, 2019.
[3] https://bioportal.bioontology.org/ontologies/SYMP; Accessed on 23rd April, 2019.

higher in the case of retrofitted embeddings than in the previous case. The highest improvement (47.55% to 51.15%) is when GloVe embeddings are retrofitted with WordNet. A minor observation is that the F-scores are lower than GloVe without the retrofitting, highlighting that retrofitting may not always result in an improvement. Table 3 shows the average performance across the seven types of word embedding initialisations. The +Pipeline approach results in a degradation of 4.78%. This shows that merely discarding tweets where the symptom word usage was predicted as figurative may not be useful. This could be because the figurative usage detection technique is not free from errors. In contrast though, for +FeatAug, there is an improvement of 2.21%. This shows that our technique of augmenting with the figurative usage-based features is beneficial. The improvement of 2.21% may seem small as compared to the prevalence of figurative tweets as described in Section 2. However, all tweets with figurative usage may not have been mis-classified by PHMD. The improvement shows that a focus on figurative usage detection helps PHMD. Finally, the F-scores for PHMD with +FeatAug with GloVe embeddings for the different illnesses, available as a part of the annotation in the dataset, is compared in Table 4. Our observation that heart attack results in the lowest F-score, is similar to the one reported in the original paper. At the same time, we observe that, except for heart attack, all illnesses witness an improvement in the case of +FeatAug.

6 Error Analysis

Typical errors made by our approach are:

- **Indirect reference**: Some tweets convey an infection by implication. For example, '*don't worry I got my face mask Charlotte, you will not catch the flu from me!*' does not specifically state that someone has influenza.

- **Health words**: In the case of stroke or heart attack, we obtain false negatives because many tweets do not contain other associated health words. Similarly, in the case of depression, some words like '*addiction*', '*mental*', '*anxiety*' appear which were not a part of the related health words taken into account.

- **Sarcasm or humour**: Some mis-classified tweets appear to be sarcastic or joking. For

example, '*I'm trying to overcome depression and I need reasons to get out the house lol*'. Here, the person is being humorous (indicated by '*lol*') but the usage of the symptom word '*depression*' is literal.

7 Related Work

Several approaches for PHM detection have been reported (Joshi et al., 2019). Lamb et al. (2013) incorporate linguistic features such as word classes, stylometry and part of speech patterns. Yin et al. (2015) use similar stylistic features like hashtags and emojis. Karisani and Agichtein (2018) implement another approach of partitioning and distorting the word embedding space to better detect PHMs, obtaining a best F-score of 69%. While we use their dataset, they use a statistical classifier while we use a deep learning-based classifier. For figurative usage detection, supervised (Liu and Hwa, 2017) as well as unsupervised (Sporleder and Li, 2009; Liu and Hwa, 2018; Muzny and Zettlemoyer, 2013; Jurgens and Pilehvar, 2015) methods have been reported. We pick the work by Liu and Hwa (2018) assuming that it is state-of-the-art.

8 Conclusions

We employed a state-of-the-art method in figurative usage detection to improve the detection of personal health mentions (PHMs) in tweets. The output of this method was combined with classifiers for detecting PHMs in two ways: (1) a simple pipeline-based approach, where the performance of PHM detection degraded; and, (2) a feature augmentation-based approach where the performance of PHM detection improved. Our observations demonstrate the promise of using figurative usage detection for PHM detection, while highlighting that a simple pipeline-based approach may not work. Other ways of combining the two modules, more sophisticated classifiers for both PHM detection and figurative usage detection, are possible directions of future work. Also, a similar application to improve disaster mention detection could be useful (for figurative sentences such as '*my heart is on fire*').

Acknowledgment

References

Lauren Charles-Smith, Tera Reynolds, Mark Cameron, Mike Conway, Eric Lau, Jennifer Olsen, Julie Pavlin, Mika Shigematsu, Laura Streichert, Katie Suda, et al. 2015. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PloS one*, 10(10):e0139701.

Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102, Melbourne, Australia. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

Antonio Jimeno Yepes, Andrew MacKinlay, and Bo Han. 2015. Investigating public health surveillance using twitter. In *Proceedings of BioNLP 15*, pages 164–170, Beijing, China. Association for Computational Linguistics.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2019. Survey of text-based epidemic intelligence: A computational linguistic perspective. *arXiv preprint arXiv:1903.05801*.

David Jurgens and Mohammad Taher Pilehvar. 2015. Reserating the awesometastic: An automatic extension of the WordNet taxonomy for novel terms. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1459–1465, Denver, Colorado. 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 World Wide Web Conference*, pages 137–146, Lyon, France.

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, Atlanta, Georgia. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3230–3236, San Francisco, CA.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731, Brussels, Belgium. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, Lake Tahoe, NV.

George A Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Conference on Artificial Intelligence*, pages 4444–4451, San Francisco, CA.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

World Health Organisation. 2019. Epidemic intelligence - systematic event detection. https://www.who.int/csr/alertresponse/epidemicintelligence/en/. [Online; accessed 24-January-2019].

Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. 2015. A scalable framework to detect personal health mentions on Twitter. *Journal of Medical Internet Research*, 17(6):e138.