A Prior Knowledge Based Neural Attention Model for Opioid Topic Identification

Riheng Yao[1,2,3] Qiudan Li[1,3] Wei-Hsuan Lo-Ciganic[4] Daniel Dajun Zeng[1,2,3]
[1]The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Shenzhen Artificial Intelligence and Data Science Institute (Longhua)
[4]Department of Pharmaceutical Outcomes & Policy, University of Florida, Gainesville, Florida, USA
{yaoriheng2017, qiudan.li}@ia.ac.cn, wlociganic@cop.ufl.edu, dajun.zeng@ia.ac.cn

***Abstract*—The opioid epidemic has become a serious public health crisis in the United States. Social media sources such as Reddit containing user-generated content may be a valuable safety surveillance platform to evaluate discussions discerning opioid use. This paper proposes a prior knowledge based neural attention model for opioid topics identification, which considers prior knowledge with attention mechanism. Experimental results on a real-world dataset show that our model can extract coherent topics, the identified less discussed but important topics provide more comprehensive information for opioid safety surveillance.**

Keywords—prior knowledge; attention; opioid; topic

I. INTRODUCTION

The epidemic of opioid use has become a serious public health crisis in the United States. In 2016, 116 Americans die from opioid-related drug overdoses every day [1]. Opioid overdose deaths involving prescription and illicit opioids more than quadrupled from 1999 to 2016 [2]. Further, the economic burden of prescription opioid misuse in the US exceeds $78 billion annually, including the costs of health care, lost productivity, treatment for substance use disorders, and criminal justice system [3]. There are urgent needs to develop or strengthen innovative surveillance systems to improve opioid safety.

Social media postings are noisy and unstructured, but provide an unprecedented opportunity to develop innovative and efficient approaches for social intelligence and health surveillance [4]. For example, one of the most popular forums in the world, Reddit, has been widely used to study a variety of issues in the health domain. On Reddit, there is a community "r/opiates" (also called "subbreddit") which includes a variety of discussion topics such as access to opioids, route of administration, and responses and reactions after opioid use. Efficiently identifying the topics of opioid-related discussions from social media may provide valuable information such as users' behaviors, preference, and utilization patterns to support health professionals, health care systems, and regulatory agencies to further improve opioid safety surveillance systems.

Prior research for topic identification are largely based on probabilistic graphical models, which do not directly capture word co-occurrences information that is primary to preserve topic coherence [5]. Address this problem, some models have integrated prior knowledge to promote the quality of topics [6]. Recently, He et al. [7] proposed a neural attention model which exploits the distribution of word-occurrences through word embedding method. Compared with topic models, this model explicitly encodes word co-occurrence information, maps words with similar semantics to be close to each other in the embedding space, thus improve the coherence of the inferred topics. However, for topics that are less discussed but are important such as sales about opioid, this method may allocate corresponding words to topics that have a large amount of discussion. If prior knowledge of less talked about topics is provided, models could be encouraged to find the evidence of these topics in corpus, thus better distinguished them from obvious ones.

This paper proposes a prior knowledge based neural attention model to identify the opioid-related topics. First, some seed words are selected for each topic as prior knowledge. Then, our model computes background representation of topics based on the word embeddings of corresponding seed words. Furthermore, the relevance of a word to a topic is measured by the cosine distance between the word embedding and the background representation. Finally, the topic embeddings are learned through sentence embedding and reconstruction process.

II. PROPOSED PRIOR KNOWLEDGE BASED MODEL

Figure 1 shows the framework of our proposed prior knowledge based neural attention model (PKBNA). The ultimate goal is to learn the embedding of each topic, then topic words will be detected according to the distance between topic embedding and word embedding. PKBNA consists of two modules: prior knowledge based word attention weight computation and sentence embedding and construction.

A. Prior Knowledge Based Word Attention Weight Computation

This module aims to obtain the attention weights of words in the sentence for sentence representation learning. Firstly, each word h in the vocabulary is associated with a representation vector $v_h \in R^n$ by neural word embedding learning, where n is the dimension of the embedding space. Then, suppose that we focus on k topics and provide some seed words of each topic. We encode the prior knowledge into a background topic representation matrix $P \in R^{k*n}$, in which the j^{th} row P_j is the average word embedding of corresponding seed words. This way of prior knowledge encoding ensures that

978-1-7281-2504-6/19/$31.00 ©2019 IEEE

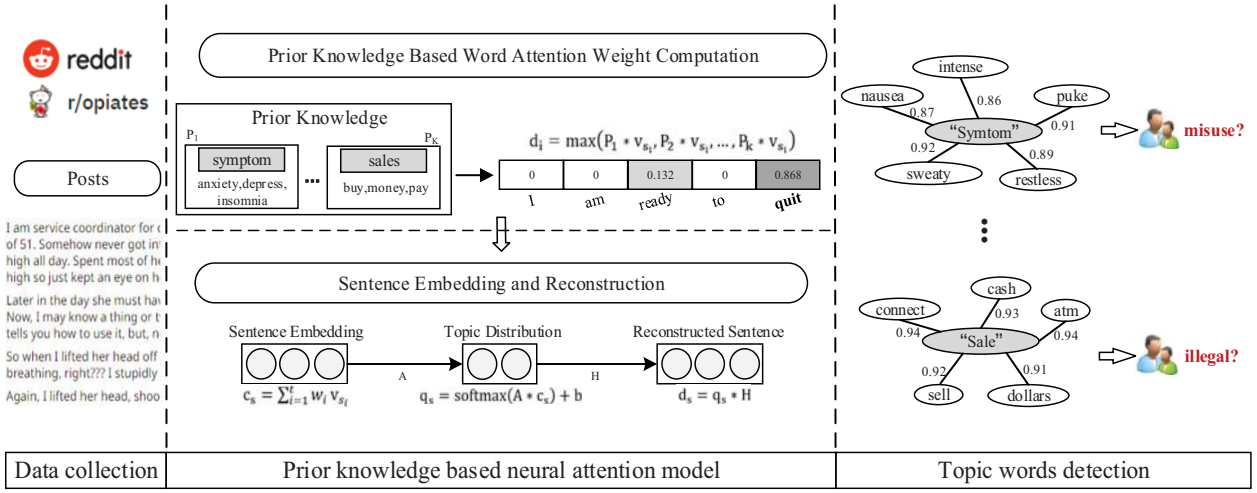

Figure 1. The framework of the proposed model

P will share the same embedding space with words. Finally, we use the background topic representation matrix to guide the model to identify prior knowledge-indicative words. Specifically, for each word s_i in a sentence that includes t words, we compute the cosine similarity between its vector v_{s_i} and each topic background representation P_j in the embedding space. The larger the cosine distance is, the more likely the word belongs to that topic, we use the maximum similarity d_i of each word to measure its attention weight w_i when modeling sentence embedding. The computation processes are formulated as follows:

$$w_i = \frac{\exp(d_i)}{\sum_l^t \exp(d_l)} \quad (1)$$

$$d_i = \max\left(P_1 * v_{s_i}, P_2 * v_{s_i}, \dots, P_k * v_{s_i}\right) \quad (2)$$

B. Sentence Embedding and Reconstruction

This module aims to learn topic embeddings $H \in R^{k*n}$, we construct sentence embeddings from weighted word embeddings and then reconstruct it as linear combination of topic embeddings from H. The embedding c_s of sentence s is calculated as the weighted summation of the embedding of words it contains, in which the weights are obtained based on the prior knowledge:

$$c_s = \sum_{i=1}^t w_i v_{s_i} \quad (3)$$

The sentence reconstruction process consists of two steps, firstly we computed the topic distributions $q_s \in R^k$ of the sentence [7]:

$$q_s = \text{softmax}(A * c_s) + b \quad (4)$$

where $A \in R^{k*n}$ and $b \in R^k$ are parameters to learn, and q_s measures the possibility that the sentence belongs to each topic. Then the reconstructed sentence representation d_s of sentence s is obtained from the linear combination from topic embeddings [7]:

$$d_s = q_s * H \quad (5)$$

To minimize the reconstruction error, the contrastive max-margin objective function [8] is used.

$$L = \sum_{s \in Y} \sum_i^e \max(0, 1 - d_s c_s + d_s m_i) \quad (6)$$

in which Y is the training corpus, e is the number of negative samples for each input sentence s and m_i is the average word embedding of sampled negative samples sentence.

III. Experimental Results and Discussions

A. Dataset

To evaluate the performance of the topic identification model, we collected data from the "r/opiates" community on Reddit between January 1, 2017 and January 1, 2018. The dataset contains 27,290 main posts and 173,552 comments, including 668,057 sentences in total.

B. Baseline Methods

- ABAE [7]: A state-of-the-art unsupervised neural attention model which captures the relevance of words to topics by taking the inner product of the word to the global context of the sentence.
- ABAE- [7]: The sentence embedding is the average embeddings of words contained.

C. Evaluation

We adopt coherence score [5] to evaluate the performance of all models. Given an topic a and its top K words, $T^a = \{w_1^a, \dots, w_k^a\}$, the coherence score is computed as formula (7), in which $F_1(w)$ is the document frequency of word w and $F_2(w_1, w_2)$ is the co-document frequency of word w_1 and w_2. Topic with higher coherence scores are better.

$$S(a; T^a) = \sum_{i=2}^K \sum_{j=1}^{i-1} log \frac{F_2\left(w_i^a, w_j^a\right)+1}{F_1(w_j^a)} \quad (7)$$

D. Experimental Settings

Stop words, punctuation symbols, numbers, emoticons, dosage like "20mg", time within a day like "3pm" and urls are removed from corpus. Also, we drop words whose frequency are less than 10.

The word embeddings are previously obtained by word2vec method [9], with embedding size of 200, window size of 5. In

Table II: Topics, seed words and some representative words only discovered by PKBNA

Topic	Seed words	Words only discovered by PKBNE
Opioids	naloxone suboxone methadone	buprenorphine diamorphine paracetamol
Symptoms	anxiety depress insomnia	sweaty puke saber
Sales	buy money pay	airport pocket downtown
Withdraw	clean detox stop	regret guilty relapse
Medical use	medic treatment prescribe	avail drugstore verify
Policy	govern policy regulate	administrator judgement statement

the training process, the number of negative samples is set to 20, we ran 15 epochs with batch size 50.

For PKBNA, we focus on 6 topics: opioids, symptoms, sales, withdraw, medical use, policy. For each topic, we select 3 seed words.

E. Analysis and Results

Table I shows the average coherence score of all topics of all models**.** It can be seen that our proposed model outperforms all baseline methods in all the ranked buckets of top words. PKBNA could well distinguish topics in all buckets, especially when the number of considered top words is larger. The good performance verifies the effectiveness of the attention mechanism based on the prior knowledge.

Table I: Average coherence score of different models.

Top N	10	20	30	40	50
PKBNA	**-214.21**	**-911.73**	**-2089.35**	**-3729.35**	**-5902.69**
ABAE	-218.29	-926.52	-2142.56	-3819.36	-6022.70
ABAE-	-214.57	-933.38	-2148.11	-3847.65	-6053.02

Table II presents the selected seed words and some representative words that only discovered by PKBNE. It can be seen that PKBNA provides more comprehensive information for opioid use analysis, which has potential to be used for practice of opioid safety surveillance. For example, through detecting more *symptoms* that are less talked about, health-care providers can better understand symptoms or adverse effects related to opioid use and develop or modify a treatment plan. The words in *sales* topic may help to target buyers and sellers of opioids, making contribution to tackle illegal sales. Identifying more words in *withdraw* topic may be applied to target opioid users at different use stages, which helps for regulators to develop different intervention plans. Finding more clues of the discussion about opioid-related *policy* could serve as feedback channel for management departments.

IV. Conclusion

In this paper, we proposed a prior knowledge based neural attention model for opioid topics identification. Under the attention guidance of prior knowledge, the model can effectively capture word-topic semantic relevance, thus better distinguish discovered topics. We also found that our model could identify more comprehensive key words of less discussed topics. Experimental results on a real world dataset verify the effectiveness of our model, which may contribute to opioid surveillance practice.

Acknowledgment

This work was supported by the National Key R&D Program of China under Grant No. 2016QY02D0305, the National Natural Science Foundation of China under Grant No. 71621002, 61671450, 71702181, the Key Research Program of the Chinese Academy of Sciences under Grant No. ZDRW-XH-2017-3.

References

[1] US Department of Health and Human Services (2018). *What is the U.S. Opioid Epidemic?* Available: https://www.hhs.gov/opioids/about-the-epidemic/index.html

[2] Centers for Disease Control and Prevention (2017). *Opioid Data Analysis* Available: https://www/cdc.gov/drugoverdose/data/analysis. html

[3] C. Florence, F. Luo, L. Xu, and C. Zhou, "The economic burden of prescription opioid overdose, abuse and dependence in the United States, 2013," *Medical care,* vol. 54, no. 10, p. 901, 2016.

[4] P. Yan, H. Chen, and D. Zeng, "Syndromic surveillance systems," *Annual review of information science and technology,* vol. 42, no. 1, pp. 425-495, 2008.

[5] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the conference on empirical methods in natural language processing*, 2011, pp. 262-272: Association for Computational Linguistics.

[6] Z. Chen, A. Mukherjee, and B. Liu, "Aspect extraction with automated prior knowledge learning," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, vol. 1, pp. 347-358.

[7] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An unsupervised neural attention model for aspect extraction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, vol. 1, pp. 388-397.

[8] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics,* vol. 2, pp. 207-218, 2014

[9] 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*, 2013, pp. 746-751.