Fine-Grained Information Identification in Health Related Posts

Hamed Khanpour
Computer Science and Eng., University of North Texas
Denton, Texas
hamedkhanpour@my.unt.edu

Cornelia Caragea
Computer Science, Kansas State University
Manhattan, Kansas
ccaragea@k-state.edu

ABSTRACT

Online health communities have become a medium for patients to share their personal experiences and interact with peers on topics related to a disease, medication, side effects, and therapeutic processes. Analyzing informational posts in these communities can provide an insightful view about the dominant health issues and can help patients find the information that they need easier. In this paper, we propose a computational model that mines user content in online health communities to detect positive experiences and suggestions on health improvement as well as negative impacts or side effects that cause suffering throughout fighting with a disease. Specifically, we combine high-level, abstract features extracted from a convolutional neural network with lexicon-based features and features extracted from a long short term memory network to capture the semantics in the data. We show that our model, with and without lexicon-based features, outperforms strong baselines.

CCS CONCEPTS

• **Information systems → Information extraction**; Retrieval;

KEYWORDS

Therapeutic processes, side effects, information extraction

ACM Reference Format:
Hamed Khanpour and Cornelia Caragea. 2018. Fine-Grained Information Identification in Health Related Posts. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3209978.3210132

1 INTRODUCTION

Traditionally, medical doctors and care providers have been the main source of information for patients who suffer from chronic or life-threatening diseases. However, with the advent of the Internet and the creation of many online health communities (OHCs), e.g., Everyday Health, Cancer Survivors' Network, and WebMD, patients use these health communities increasingly as an integral source for finding health-related information [5]. OHCs provide an environment for patients, their family members and friends to interact with other participants and share experiences and information (e.g., recommendations and feedback) on issues related to prescribed medicines, side effects, therapeutic processes, mental health, and feelings. Table 1 shows examples of posts that contain health-related information shared among patients in an online cancer community. This information is very unique and is often not available elsewhere, e.g., referring to the medication Sertraline, a patient writes: *Doctors also say it helps with hot flashes. I don't know about that since I still get them* (see Example 1 in the table).

Exp. 1: I am on Sertraline, which is generic zoloft and I truly believe it has helped me. I have been on it since I was initially diagnosed. I took it all thru chemo and I am still on it. Doctors also say it helps with hot flashes. I don't know about that since I still get them. But at least I am not depressed. So that is good.
Exp. 2: I took Anzamet... one pill prior to the infusions and one each day for 3 days following treatments. The only real problem I developed with it, and it lasted till I finished treatment, was an aversion to drinking water! Plain water began to just taste terrible to me.

Table 1: Examples of OHC informational messages.

Several studies showed that using OHCs to obtain information from people who went through the same or similar experiences (either by direct interactions or sifting through the online posts) brings better feelings and fewer mortality odds to patients [8]. Thus, the large and growing amounts of user-generated content in OHCs need to be accurately classified for a variety of applications, e.g., designing smart information retrieval systems for content recommendation. Recent computational studies in OHCs started to investigate the high level identification of informational posts [1, 19], however, with no emphasis on the unique challenges associated with the detection of the information type, e.g., therapeutic procedures vs. side effects. A deep understanding of the text and the writer's intention is required in order to correctly extract the types of information present in OHCs messages. Example 1 in Table 1 refers to therapeutic procedure, whereas Example 2 refers to side effects through various medication (Sertraline and Anzamet, respectively).

In this paper, we propose to analyze messages in OHCs to extract the information type that they contain, i.e., *therapeutic procedures* (any medical treatment, activity, or behavior that have a positive impact on patients' health, precisely, can help prevent, cure or improve a patient's condition) and *side effects* (any medical treatment, activity, or behavior that have a negative impact on patients' health, precisely, a secondary, often undesirable effect of a drug or medical treatment). To achieve this, we design a computational model that is able to exploit the semantic information from text, and coherently combines high-level (abstract) features with surface-level and lexicon-based features. Our contributions are as follows:

(1) We propose to extract fine-grained information types from messages posted in OHCs. Identifying information types provides doctors, health practitioners and OHCs' moderators with an insightful view of patients' physical status during

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210132

various treatments. In addition, it can provide new diagnosed patients with information about what they should expect throughout their treatments and help them in making informed decisions about their disease more effectively [14]. To our knowledge, we are the first to address fine-grained information type extraction in OHCs.

(2) We design and explore a computational model that can identify messages belonging to *therapeutic procedures* and *side effects* with high accuracy. Our model is a hybrid neural network combined with lexicon-based features.

(3) We show empirically that our model significantly outperforms strong baselines and prior works and continues to perform well even in the absence of lexicon-based features.

2 RELATED WORK

In computational studies, messages in OHCs have been analyzed from the standpoint of social support [3, 7], with emotional [5] and informational [2] support being the two principle functions that shape the majority of messages in OHCs. Thus far, most computational studies in OHCs are dedicated to analyzing and identifying messages that contain these two types of support. For example, Wang et al. [19] used a linear regression model to identify emotional and informational support in messages from a cancer forum and studied the relationship of these support types on user engagement with the health community. Their feature set includes: LIWC features, POS tags, message length, subjectivity intensity, and Latent Dirichlet Allocation based topical features.

Biyani et al. [1] learned classifiers (e.g., Naïve Bayes and Logistic Regression) to classify messages that contain emotional or informational support from posts in a breast cancer discussion board of a cancer survivors' network. The authors used unigrams, POS tags, structural linguistic patterns, and five lexicons that contain strong and weak subjective words, cancer drugs, side-effects, and cancer procedures, and showed that features drawn from lexicons have the highest impact on the results. On a breast cancer dataset constructed from the Cancer Survivors' Network (CSN) of the American Cancer Society (ACS), the authors showed that their classifiers can identify emotional and informational messages with an F1-score of 0.88 and 0.77, respectively. Furthermore, Wang et al. [18] studied the correlation between social support and user engagement, but instead of using a regression model as in [19], the authors used traditional machine learning classifiers such as Naïve Bayes, Logistic Regression, Support Vector Machines, Random Forest to classify OHCs' messages based on the intention of the participant when writing a message (i.e., companionship, seeking information, seeking emotion, providing information, and providing emotion). The authors used a combination of features from Wang et al. [19] coupled with lexicon-based features used in Biyani et al. [1]. Sondhi et al. [15] extracted sentences from medical forums that contain medical problems and medical treatments, using Support Vector Machines and Conditional Random Fields, trained on novel features such as semantic features, position based features, and user based features and achieved an accuracy as high as 75%.

In contrast to the above works that mainly used traditional machine learning, we focus on the unique challenges associated with fine-grained detection of informational messages, i.e., messages belonging to the categories *therapeutic procedures* and *side effects*, using a hybrid deep neural network. Our task has the potential to improve patients' competence and knowledge in dealing with health care problems and will empower them to become better prepared and take control of their life in better ways.

3 DATA COLLECTION AND ANNOTATION

Since there is no available dataset for analyzing messages that contain fine-grained informational content in OHCs (i.e., therapeutic procedures and side effects), we constructed a benchmark dataset to evaluate our model. We randomly selected 225 comments from 21 discussion threads in the lung cancer discussion board and 120 comments from 11 discussion threads in the prostate cancer discussion board of CSN. We performed our data annotation at sentence level and selected sentences with length greater than four words to exclude appreciative and appraisal messages. We obtained 1,797 sentences, which were integrated with the 1,066 sentences extracted from the breast cancer discussion board in CSN that were provided by Biyani et al. 2014, with an overall 2,863 sentences.

The purpose of the annotation was to label the 2,863 sentences as belonging to *therapeutic procedures*, i.e., containing information about any medical treatment, activity, or behavior that have a positive impact on patients' health; *side effects*, i.e., containing information about any medical treatment, activity, or behavior that have a negative impact on patients' health; and *other*, which includes sentences that do not belong to any of the above two categories. Two annotators (graduate students) contributed to the task, which was conducted in an iterative fashion following prior studies and guidelines [4, 6, 13]. In each round, 200 messages were assigned and annotators discussed disagreements with researchers; 100% inter-annotator agreement (IAA) was achieved after each round of discussions. We used Cohen's kappa for measuring IAA. After four initial rounds, the remaining data (2,063 messages) were assigned to the annotators where they achieved 90% IAA. The last round of the assigned data was adjudicated by one of the authors.

Table 2 provides the distribution of messages in each category. As can be seen, *therapeutic procedures* has significantly more sentences than *side effects*. This shows that patients tend to share more of their success stories and positive aspects of their therapy rather than sharing negative impacts or side-effects. The category *other* has the largest number of sentences. A large fraction of these sentences contain emotional support such as empathy and encouragements.

Category	#sentences	percentage (%)
therapeutic procedures	942	32.9
side effects	385	13.4
other	1536	53.7

Table 2: Statistics from the data collection.

4 MODEL

In this section, we describe our proposed computational model, which can be embedded in OHCs' search engines to retrieve fine-grained messages belonging to *therapeutic procedures* or *side effects*. Given a sentence of n words, we apply Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) networks concurrently. The CNN extracts high-level (abstract) features that capture the semantic part of the text [10], whereas the LSTM captures sequential information from each sentence.

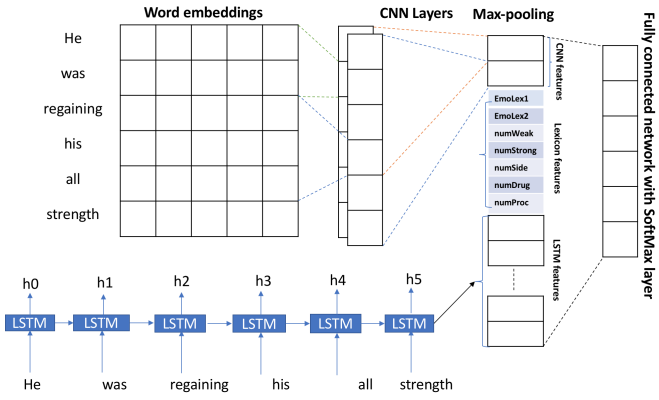

Figure 1: The architecture of our proposed Hybrid Neural Network with Lexicon-based features (HNNL).

Our CNN architecture consists of one convolution layer followed by a max pooling layer. The input data layer is fed with word vectors of length k, where $x_i \in \mathbb{R}^k$ is the k-dimensional word vector corresponding to the i-th word in the sentence. Thus, the input sentence of length n is represented as

$$x_{1:n} = x_1 \oplus x_2 \oplus .. \oplus x_i \oplus ... \oplus x_n \quad (1)$$

where $\oplus$ is the concatenation operator. First, l filters for each region size (rs) are applied to the sequence of tokens in each sentence (e.g., 16 filters for 2 and 3 region sizes). The feature map M belongs to $\mathbb{R}^{l \times rs}$. The features m_j, with $j = 1, \cdots, T$ (where T is the number of extracted features), are defined as follows:

$$m_j = Relu(M[x_{j-(rs/2)+1}, \cdots, x_j, \cdots, x_{j+(rs/2)}]) \quad (2)$$

where $Relu$ is the rectified linear unit activation function. This process is iteratively done for each time step (corresponding to each word) of the input sentence that ends up with $M = (m_1, m_2, \cdots, m_T)$ sequence. Second, max pooling is applied to M, which results in $M' = (m'_1, m'_2, ..., m'_{T/(pooling-size)})$. M' is the output of the CNN that contains high-level, abstract features.

The LSTM unit consists of sub-unit-inputs (i_t), output (o_t), forget gates (f_t) and memory cell (c_t).

$$i_t = \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}\right) \quad (3)$$

$$f_t = \sigma\left(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}\right) \quad (4)$$

$$o_t = \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}\right) \quad (5)$$

$$u_t = \tanh\left(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}\right) \quad (6)$$

LSTM unit at time t computes the memory cell:

$$u_t = \tanh(Wx_t + Uh_{t-1} + b) \quad (7)$$

$$c_t = i_t \bigodot u_t + f_{(t)} \bigodot c_{t-1} \quad (8)$$

and then computes the output, or activation:

$$h_t = o_t \bigodot \tanh(c_t) \quad (9)$$

Here, $x \in \mathbb{R}^{n \times k}$ is the input and $W \in \mathbb{R}^{n \times k}$, $U \in \mathbb{R}^{n \times n}$, and $b \in \mathbb{R}^n$ are parameters of an affine transformation. The resulting sequence of the layers is $h_1, h_2, \cdots, h_n$.

Last, we combine the features extracted by CNN and LSTM networks with lexicon-based features in a hybrid model. Our proposed model, HNNL, is shown graphically in Figure 1. As can be seen from the figure, we use a combination of CNN and LSTM models, where the final feature vectors from CNN augmented with lexicon-based features and the last feature h_n extracted by LSTM are fed into a fully connected network with a SoftMax layer.

HNNL: LR= 0.1, Decay rate=0.6, Dropout=0.8, Layer=1, Max pooling, FRS=(2,3), NF=16
HNN: LR= 0.1, Decay rate=0.6, Dropout=0.8, Layer=1, Max pooling, FRS=(2,3), NF=16
LSTM: LR= 0.001, L2reg=$1E-5$, Decay rate=0.7, Layer=1, Max pooling
CNN: LR= 0.1, Decay rate=0.5, Dropout=0.6, Layer=2, Max pooling, FRS=(2,3,4), NF=16
ConvLSTM: LR= 0.1, Decay rate=0.7, Dropout=0.6, Layer=2, Max pooling, FRS=(2,3), NF=16
ConvLexLSTM: Decay rate=0.8, Dropout=0.5, Layer=1, Max pooling, FRS=(2,3), NF=16

Table 3: Hyper-parameter settings for all models.

Lexicon-based Features: In this work, we used seven lexicons. The first five lexicons come from Biyani et al. 2014. These lexicons are: weak subjective words (**numWeak**), strong subjective words (**numStrong**) cancer drugs (**numDrug**), side-effects (**numSide**), and therapeutic procedures (**numProc**). The sixth and seventh lexicons come from emotion detection research. Our motivation for the integration of these two emotion lexicons is that a large fraction of sentences in the category *other* are emotional in nature, where people emotionally support one another. These lexicons are: **EmoLex1** [16] and **EmoLex2** [11]. We use frequencies of lexicon words to construct the lexicon-based features.

5 EXPERIMENTS AND RESULTS

Next, we describe the evaluation of HNNL using binary tasks. Specifically, we trained our models in the two-class setting by binarizing the datasets: *therapeutic procedures* vs. *non-therapeutic procedures* (and *side effects* vs. *non-side effects*). In all experiments, we used word embeddings as input to the neural networks, which were generated with the W2vector module in Gensim [12] on the data from all discussion boards of CSN between 2000 and 2012. The results (weighted average Precision, Recall and F1-score) are reported in 10-fold cross validation experiments.

Hyper-parameter setting: We optimized hyper-parameter values of our HNNL model as well as all the other neural network models (used for comparison) by performing a grid search on a development set, which consists of 20% of instances removed from the training set in each iteration of 10-fold cross-validation. Table 3 shows the best hyper-parameter values for all neural network models.

Performance of HNNL in an ablation experiment. Since our HNNL model is a hybrid neural network model with several components, first, we evaluate its performance in an ablation experiment to understand the contribution of each component in the model performance. Specifically, we compare HNNL with HNN (a model that has the same architecture as HNNL, but does not use any external lexicon), CNN, LSTM, and support vector machines (SVM) with the (concatenated) features from the seven lexicons (described above).

Table 4 shows the results of these comparisons (first block of results). As can be seen, HNNL achieves the best results consistently throughout all experiments in terms of all compared measures. This ablation experiment confirms that all components in our model positively contribute to the final results. For example, eliminating

Method	TP			SE		
	Pr	**Re**	**F1**	**Pr**	**Re**	**F1**
HNNL	**82.9**	**87.1**	**84.9**	**79.8**	**81.3**	**80.5**
HNN	80.3	84.6	82.3	77.5	79.3	78.4
LSTM	75.8	78.2	76.9	71.2	73.0	72.0
CNN	70.0	72.3	71.1	68.0	69.7	68.8
Seven-Lexicon	69.2	65.0	67.0	65.2	67.1	66.1
C-ConvLSTM	78.1	75.6	76.8	76.6	71.0	73.6
LibShortText toolkit6	69.9	72.4	71.1	68.5	69.6	69.0
Tf-Idf	62.7	63.8	63.2	60.1	63.0	61.5
ConvLexLSTM	79.5	82.9	81.1	76.4	77.7	77.0
ConvLSTM	78.6	79.9	79.2	73.7	77.0	75.3

Table 4: Classification results of HNNL vs. other models. TP denotes *therapeutic procedures*, and SE denotes *side effects*.

the seven lexicon features from HNNL, which yields HNN, results in a drop in F1-score by 2.6% on *therapeutic procedures* and by 2.1% on *side effects* classification results. Still, HNN is the second performing model in terms of F1-score. These results show that our model can be successfully applied in a health domain even in the absence of health lexicons, which are often expensive to obtain and require domain experts to design them. Not surprisingly, the SVM with the seven lexicon-based features (denoted as Seven-Lexicon) performs the worst among the compared models, suggesting that obtaining the semantic information from text improves models' performance.

Baseline Comparisons. Second, we compare HNNL with several baselines and prior works: (1) C-ConvLSTM (i.e., a character-level CNN-LSTM) by Kim et al. [9] in which the output of CNN is input for LSTM; (2) LibShortText toolkit6 that uses SVM with part of speech tags and other syntactic and semantic features such as frame-semantics, dependency triples, and (3) an SVM with *tf-idf* features. The LibShortText toolkit6 has been shown to have a very good performance on classifying short-texts [17]. Table 4 shows the results of this comparison in the second block of results. As can be seen, HNNL and HNN outperform all three baselines, and more importantly, they outperform the C-ConvLSTM, which represents a different (i.e., serial) combination of CNN and LSTM networks, however, at character level. Interestingly, how would the HNNL model that uses CNN and LSTM in parallel compare with a (serial) combination of CNN and LSTM at word level? To understand this, we designed a model called ConvLSTM, i.e., a word-level CNN-LSTM that uses word embeddings trained on CSN data instead of character-level CNN-LSTM as in C-ConvLSTM. We also extended ConvLSTM to include the seven lexicons into the ConvLexLSTM model. More precisely, in ConvLexLSTM, the CNN features augmented with lexicon-based features are fed as input to LSTM. The results of these two baselines are shown in the last block of Table 4. As can be seen from the table, the performance of word-level ConvLSTM is higher than character-level C-ConvLSTM. Adding the lexicon features to ConvLSTM yields even higher performance, but not as high as that of HNNL. This result confirms our belief that preserving the sequential information in sentences added to the CNN features yields improvement in performance over models which do not include sequential information.

It is also worth mentioning that all deep neural networks that capture semantics from the data perform better that the traditional models. The lexicon-based features act as a complement (for the high-level semantic features) by finding exact words in the text to generate proper features in HNNL.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed a computational model for classifying fine-grained informational messages in OHCs. Our proposed model, HNNL, combines the strengths of CNNs, LSTMs, and lexicon-based approaches to capture hidden semantics in OHCs' messages. We show that our proposed model, with or without lexicon-based features, which are often expensive to obtain or maintain in a health domain, provides a better computational model for classifying informational messages based on their content compared with strong baselines, including other types of deep neural networks. In future, it would be interesting to study the performance of our models on data from different health communities, e.g., related to weight loss.

ACKNOWLEDGMENTS

We are grateful to the American Cancer Society for making the Cancer Survivors' Network available to us. We thank Iulia Bivolaru and Manoj Panchagnula for their help with data annotation.

REFERENCES

[1] Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying Emotional and Informational Support in Online Health Communities.. In *COLING*. 827–836.
[2] Heather S Boon, Folashade Olatunde, and Suzanna M Zick. 2007. Trends in complementary/alternative medicine use by breast cancer survivors: comparing survey data from 1998 and 2005. *BMC women's health* 7, 1 (2007), 4.
[3] Cindy-Lee Dennis. 2003. Peer support within a health care context: a concept analysis. *International journal of nursing studies* 40, 3 (2003), 321–332.
[4] Sidney K D'Mello. 2016. On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing* 7, 2 (2016), 136–149.
[5] Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ* (2004).
[6] Karën Fort et al. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Wiley Online Library.
[7] Howard S Friedman and Roxane Cohen Silver. 2007. *Foundations of health psychology*. Oxford University Press.
[8] Julianne Holt-Lunstad, Timothy B Smith, and J Bradley Layton. 2010. Social relationships and mortality risk: a meta-analytic review. *PLoS Med* 7, 7 (2010).
[9] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-Aware Neural Language Models.. In *AAAI*. 2741–2749.
[10] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification.. In *AAAI*, Vol. 333. 2267–2273.
[11] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.
[12] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50.
[13] James G Shanahan, Yan Qu, and Janyce Wiebe. 2006. *Computing attitude and affect in text: Theory and applications*. Vol. 20. Springer.
[14] AH Shennan and S Bewley. 2005. Are virtual communities good for our health? *birth* 90 (2005), F134–40.
[15] Parikshit Sondhi, Manish Gupta, ChengXiang Zhai, and Julia Hockenmaier. 2010. Shallow Information Extraction from Medical Forum Data. In *Proceedings of COLING '10*. 1158–1166.
[16] Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *4th Semantic Evaluations*. 70–74.
[17] William Yang Wang and Diyi Yang. [n. d.]. That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using# petpeeve Tweets.
[18] Xi Wang, Kang Zhao, and Nick Street. 2014. Social support and user engagement in online health communities. In *Smart Health*. Springer, 97–110.
[19] Yi-Chia Wang, Robert Kraut, and John M Levine. 2012. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *CSCW*. ACM, 833–842.