Sentiment-Aware Recommendation System for Healthcare using Social Media

Alan Aipe, Mukuntha N S, Asif Ekbal

Department of Computer Science and Engineering
Indian Institute of Technology Patna
Patna, India
{alan.me14,mukuntha.cs16,asif}@iitp.ac.in

Abstract. Over the last decade, health communities (known as forums) have evolved into platforms where more and more users share their medical experiences, thereby seeking guidance and interacting with people of the community. The shared content, though informal and unstructured in nature, contains valuable medical and/or health related information and can be leveraged to produce structured suggestions to the common people. In this paper, at first we propose a stacked deep learning model for sentiment analysis from the medical forum data. The stacked model comprises of Convolutional Neural Network (CNN) followed by a Long Short Term Memory (LSTM) and then by another CNN. For a blog classified with positive sentiment, we retrieve the top-n similar posts. Thereafter, we develop a probabilistic model for suggesting the suitable treatments or procedures for a particular disease or health condition. We believe that integration of medical sentiment and suggestion would be beneficial to the users for finding the relevant contents regarding medications and medical conditions, without having to manually stroll through a large amount of unstructured contents.

Keywords: Health social media · Deep learning · Suggestion mining · Medical sentiment

1 Introduction

With the increasing popularity of electronic-bulletin boards, there has been a phenomenal growth in the amount of social media information available online. Users post about their experiences on social media such as medical forums and message-boards, seeking guidance and emotional support from the online community. As discussed in [2], medical social media is an increasingly viable source of useful information. These users, who are often patients themselves or the friends and/or relatives of patients write their personal views and/or experiences. Their posts are rich in information such as their experiences with disease and their satisfaction with treatment methods and diagnosis.

As discussed in [3], medical sentiment refers to a patient's health status, medical conditions and treatment. Extraction of this information as well as its analysis have several potential applications. The difficulty in the extraction of information such as sentiment and suggestions from a forum post can be attributed to a variety of reasons. Forum posts contain informal language combined with the usages of medical conditions and terms. The medical domain itself is sensitive to misinformation. Thus, any system built on this data would also have to incorporate relevant domain knowledge.

1.1 Problem Definition

Our main objective is to develop a sentiment-aware recommendation system to help build a patient assisted health-care system. We propose a novel framework for mining medical sentiments and suggestions from medical forum data. This broad objective can be modularized into the following set of research questions:

RQ1: Can an efficient multi-class classifier be developed to help us understand the overall medical sentiment expressed in a medical forum post?

RQ2: How can we model the similarity between the two medical forum posts?

RQ3: Can we propose an effective algorithm for treatment suggestion by leveraging medical sentiment obtained from the forum posts?

By addressing these research questions, we aim to create a patient assisted health-care system, which is able to determine the sentiments of any user that s/he expresses in the forum post, point the user to similar forum posts for more information, and suggest possible treatment(s) or procedural methods for the user's symptoms and possible disorders.

1.2 Motivation

The amount of health related information being sought after on the Internet is on the rise. As discussed in [4], an estimated 6.75 million health-related searches are made on Google every day. The Pew Internet Survey [5] claims that 35% of U.S. adults have used the internet to diagnose a medical condition they themselves or another might have, and that 41% of these online diagnosers have had their suspicions confirmed by a clinician. There has also been an increase in the number of health-related forums and discussion boards on the Internet, which contain useful information that is yet to be properly harnessed. Sentiment analysis has various applications. We believe it can also provide important information in health-care. In addition to doctor's advice, connecting with other people who have been in similar situations can help with several practical difficulties. According to The Pew Internet Survey, 24% of all adults have obtained information or support from others who are having the same health conditions.

A person posting on such a forum is often looking for emotional support from similar people. Consider the following two blog posts:

> **Post 1**: Hi. I have been on sertaking 50mgs for about 2 months now and previously was at 25mg for two weeks. Overall my mood is alot more stable and I dont worry as much as I did before however I thought I would have a bath and when I dried my hair etc I started to feel anxious, lightheaded and all the lovely feeling you get with panic. Jus feel so yuck at the moment but my day was actually fine. This one just came out of the blue.. I wanted to know if anyone else still gets some bad moments on these. I don't know if they feel more intense as I have been feeling good for a while now. Would love to hear others stories.

> **Post 2**: Just wanna let you all know who are suffering with head aches/pressure that I finally went to the doctor. Told him how mines been lasting close to 6 weeks and he did a routine check up and says he's pretty I have chronic tension headaches. He prescribed me muscle relaxers, 6 visits to neck massages at a physical therapist and told me some neck exercises to do. I went in on Tuesday and since yesterday morning things have gotten better. I'm so happy I'm finally getting my life back. Just wanted you all to know so maybe you can feel better soon

In the first post, the author discusses an experience with a drug, and is looking to hear from people with similar issues. In the second post, the author discusses a positive experience and seeks to help people with similar problems. One of our aims is to develop a system to automatically retrieve and provide such a user with the posts most similar to theirs. Also, in order to make an informed decision knowing patient's satisfaction for a given course of treatment might be useful. We also seek to provide suggestions for treatment for a particular patient's problems. The suggestions can subsequently be verified by a qualified professional, and then be prescribed to the patients, or in more innocuous cases (such as with 'more sleep' or 'meditation'), can be directly taken as advice.

1.3 Contributions

In this paper, we propose a sentiment-aware patient assisted health-care system using the information extracted from the medical forums. We propose a deep learning model with a stacked architecture that makes use of Convolutional Neural Network (CNN) layers, and a Long-Short Term Memory (LSTM) network, for the classification of a blog post into its medical sentiment class. To the best of our knowledge, exploring the usage of medical sentiment to retrieve similar posts (from medical blogs) and treatment options is not yet attempted. We summarize the contributions of our proposed work as follows:

- We propose an effective deep learning based stacked model utilizing CNN and LSTM for medical sentiment classification.

- We develop a method for retrieving the relevant medical forum posts, similar to a given post.
- We propose an effective algorithm for treatment suggestions, that could lead towards building a patient care system.

2 Related Works

Social media is a source of huge information that can be leveraged for building many socially intelligent systems. Sentiment analysis has been explored quite extensively in various domains. However, this has not been addressed in the medical/health domain in the required measure.

In [3], authors have analyzed the peculiarities of sentiment and word usage in medical forums, and performed quantitative analysis on clinical narratives and medical social media resources. In [2], multiple notions of sentiment analysis with reference to medical text are discussed in details. In [1], authors have built a system that identifies drugs that cause serious adverse reactions, using messages discussing them from online health forums. They use an ensemble of Naïve Bayes (NB) and Support Vector Machines (SVMs) classifiers to successfully identify the past drugs withdrawn from the market. Similarly, in [13], users' written contents from social media were used to mine the association between drugs for predicting the Adverse Drug Reactions (ADRs). FDA alerts were used as gold standard, and the statistic Proportional Reporting Ratios (PRR) was shown to be of high importance in solving the problem. In [11], one of the shared tasks involved the retrieval of medical forum posts related to the search queries provided. The queries involved were short, detailed and to the point, typically being less than 10 words. Our work however, focuses more on medical sentiment involved in an entire forum post, and helps to retrieve the similar posts. Recently, [12] presented a benchmark setup for analyzing the medical sentiment of users on social media. They identified and analyzed multiple forms of medical sentiments in text from forum posts, and developed a corresponding annotation scheme. They also annotated and released a benchmark dataset for the problem.

In our current work we propose a novel stacked deep learning based ensemble model for sentiment analysis in the medical domain. This is significantly different from the prior works mentioned above. To the best of our knowledge, no prior attempt has been made to exploit medical sentiment from social media posts to suggest treatment options to build a patient-assisted recommendation system.

3 Proposed Framework

In this section, we describe our proposed framework comprising of three phases, each of which tackles a research question enumerated in Section 1.1.

3.1 Sentiment Classification

Medical sentiment refers to analyzing the health status reflected in a given social media post. We approach this task as a multi-class classification problem using sentiment taxonomy as described in Section 4.1. Convolutional Neural Network (CNN) architectures have been extensively applied to sentiment analysis and classification tasks [8,12]. Long Short Term Memory (LSTMs) are a special kind of Recurrent Neural Network (RNN) capable of learning long-term dependencies by handling the vanishing and exploding gradient problem [6]. We propose an architecture consisting of two deep Convolutional Neural Network (CNN) layers and a Long-Short Term Memory (LSTM) layer, stacked with a fully connected layer followed by three-neurons output layer having softmax as an activation function. A diagrammatic representation of the classifier is shown in Fig. 1. The social media posts are first vectorized (as discussed in Section 4.3) and then fed as input to the classifier. Convolutional layers, used in the classifier, generate 200 dimensional feature maps of unigram and bigram filter sizes. Feature maps from the final CNN layer are maxpooled, flattened and fed into the fully connected layer having a rectified linear unit (ReLU) activation function. The output of the above mentioned layer is fed into another fully connected layer with a softmax activation to obtain class probabilities. Sentiment denoted by the class having the highest softmax value is considered to be the medical sentiment of the input message. The intuition behind adopting a CNN-LSTM-CNN architecture is as follows: During close scrutiny of the dataset, we observed that users often share experiences adhering to their time frame. For example, “I was suffering from anxiety. My doctor asked me to take cit 20mg per day. Now I feel better”. In this post, the user portrays his/her initial condition, explains the treatment which was used and then the effect of the treatment- all in a very timely sequence. Moreover, health status also keeps changing in the same sequence. This trend was observed throughout the dataset. Therefore, temporal features are the key to medical sentiment classification. Hence, in our stacked ensemble model, first CNN layer extracts top-level features, then LSTM finds the temporal relationships between the extracted features and the final CNN layer filters out the top temporal relationships which are subsequently fed into a fully connected layer.

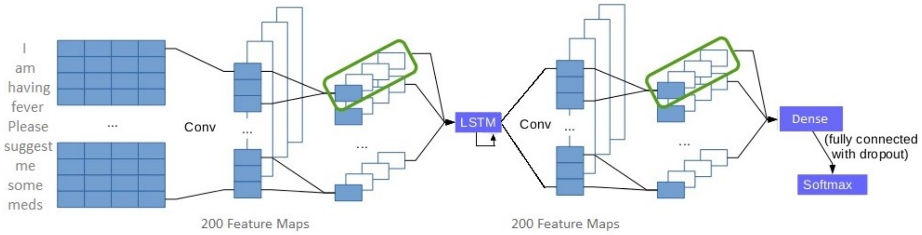


Fig. 1. Stacked CNN-LSTM-CNN Ensemble Architecture for medical sentiment classification

3.2 Top-N Similar Posts Retrieval

Users often share contents in forums, seeking guidance and to connect with other people who experienced similar medical scenarios. Thus, retrieving top-N similar posts would help to focus on contents which are relevant to one’s medical condition, without having to manually scan through all the forum posts. We could have posed this task as a regression problem where a machine learning

(ML) model learns to predict similarity score for a given pair of forum posts, but there is no suitable dataset available for this task to the best of our knowledge. We tackle this task by creating a similarity metric (as shown in E.q. 5) and evaluating it by manually annotating a small test set (as discussed in Section 4.5). The similarity metric comprises of three terms:

- *Disease Similarity*: It refers to the Jaccard similarity score computed between the two forum posts with respect to the diseases mentioned in the posts. Section 4.4 discusses how the diseases are extracted from a given post. Let *J(A,B)* denotes the Jaccard similarity between set A and B, *DS(P,Q)* denotes the disease similarity between two forum posts P and Q, *D(P)* and *D(Q)* denote the set of diseases mentioned in P and Q respectively, then:

$$DS(P,Q) = J(D(P), D(Q)) \tag{1}$$

  where,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- *Symptom similarity* : It refers to the Jaccard similarity between two forum posts with respect to the symptoms mentioned in them. Section 4.4 discusses how the symptoms mentioned in texts are extracted from a given post. Let *SS(P,Q)* denotes the disease similarity between two forum posts P and Q, *S(P)* and *S(Q)* denote the set of diseases mentioned in P and Q respectively, then

$$SS(P,Q) = J(S(P), S(Q)) \tag{2}$$

- *Text similarity* : It refers to the cosine similarity between the document vectors corresponding to two forum posts. Document vector of a post is the sum of vectors of all the words (Section 4.3) in a given sentence. Let $\boldsymbol{D_P}$ and $\boldsymbol{D_Q}$ denote the document vectors corresponding to the forum posts P and Q, *TS(P,Q)* denotes the cosine similarity between them, then

$$TS(P,Q) = \frac{\boldsymbol{D_P} \cdot \boldsymbol{D_Q}}{|\boldsymbol{D_P}| \times |\boldsymbol{D_Q}|} \tag{3}$$

We compute the above similarities between a pair of posts, and use Equation 5 to obtain the overall similarity score *Sim(P,Q)* between two given forum posts P and Q. For a given test instance, training posts are ranked according to the similarity score (with respect to test) and top-N posts are retrieved.

$$MISim(P,Q) = \frac{2 \times DS(P,Q) + SS(P,Q)}{3} \tag{4}$$

$$Sim(P,Q) = \frac{2 \times MISim(P,Q) + TS(P,Q)}{3} \tag{5}$$

where *MISim(P,Q)* denotes the similarity between P and Q with respect to the relevant medical information.

The main objective of similar post retrieval is to search for the posts depicting similar medical experience. Medical information shared in a forum post can be considered as an aggregate of the disease conditions and symptoms encountered. Medical experience shared in a forum post can be considered as an aggregate of the medical information shared and the semantic meaning of the text, in the same order of relevance. This is the intuition behind adoption of the similarity metric in Equation 5.

3.3 Treatment Suggestion

A treatment T mentioned in a forum post P can be considered suitable for a disease D mentioned in post Q if P and Q depict similar medical experience and the probability that T to produce a positive medical sentiment, given D. Thus, suggestion score *G(T,D)* is given by,

$$G(T,D) = Sim(P,Q) \times Pr(+veSentiment|T,D) \tag{6}$$

$$\begin{aligned} G(T,D) &\geq \tau \qquad && \text{Treatment T is suggested} \\ &< \tau \qquad && \text{Treatment T is not suggested} \end{aligned}$$

where τ is a hyper-parameter of the framework and *Pr(A)* denotes the probability of event A.

4 Dataset and Experimental Setup

In this section, we discuss the details of the datasets used for our experiments and the evaluation setups.

4.1 Forum dataset

We perform experiments using a recently released datasets for sentiment analysis [12]. This dataset consists of social media posts collected from the medical forum 'patient.info'. In total 5,189 posts were segregated into three classes – Exist, Recover, Deteriorate based on medical conditions the post described, and 3,675 posts were classified into three classes – Effective, Ineffective, Serious Adverse Effect based on the effect of medication. As our framework operates at a generic level, we combine both the segments into a single dataset, mapping labels from each segment to a sentiment taxonomy as discussed in Section 4.1. The classes with respect to medical condition are redefined as follows:

- **Exist**: User shares the symptoms of any medical problem. This is mapped to the *neutral sentiment.*
- **Recover**: Users share their recovery status from the previous problems. This is mapped to the *positive sentiment.*
- **Deteriorate**: User share information about their worsening health conditions. We map this to the *negative sentiment.*

The classes with respect to the effect of medication are:

Sentiment	Distribution(%)
Positive	37.49
Neutral	32.34
Negative	30.17

Table 1. Class distribution in the dataset with respect to sentiment taxonomy

- **Effective**: User shares information about the usefulness of treatment. This is mapped to the *positive sentiment*.
- **Ineffective**: User shares information that the treatment undergone has no effect as such. These are mapped to the *neutral sentiment*.
- **Serious adverse effect**: User shares negative opinions towards the treatment, mainly due to adverse drug effect. This is mapped to the *negative sentiment*.

Sentiment Taxonomy A different sentiment taxonomy is conceptualized keeping in mind the generic behavior of our proposed system. It does not distinguish between the forum posts related to medical conditions and medication. Thus, a one-to-one mapping from sentiment classes used in each segment of the dataset to a more generic taxonomy is essential. We show the class distribution in Table 1.

- **Positive sentiment** : Forum posts depicting improvement in overall health status or positive results of the treatment.
  For example : ”I have been suffering from anxiety for a couple of years. Yesterday, my doc prescribed me Xanax. I am feeling better now.” This post is considered positive as it depicts positive results of Xanax.
- **Negative sentiment** : Forum posts describing deteriorating health status or negative results of treatment.
  For example : ”Can citalopram make it really hard for you to sleep? i cant sleep i feel wide awake every night for the last week and im on it for 7 weeks.”
- **Neutral sentiment**: This denotes to the forum posts where neither positive nor negative sentiment is expressed, with no change in overall health status of the person.
  For example : ”I was wondering if anyone has used Xanax for anxiety and stress. I have a doctors appointment tomorrow and not sure what will be decided to use.”

4.2 Word Embeddings

Capturing semantic similarity between the target texts is an important step towards accurate classification. For this reason, word embeddings play a pivotal role. We use the pre-trained word2vec [10] model[1], induced from the PubMed and PMC texts along with the texts extracted from the Wikipedia dump.

[1] http://bio.nlplab.org/

4.3 Tools used and Preprocessing

The codebase, during experimentation, is written in Python (version 3.6) with external libraries – namely *keras*[2] for neural network design, *sklearn*[3] for evaluation of baseline and the proposed model, *pandas*[4] for easier access of data in the form of tables (or, data frames) during execution, *nltk*[5] for textual analysis and *pickle*[6] for saving and retrieving input-output of different modules from the secondary storage devices. The preprocessing phase comprises of the removal of non-ASCII characters, stop words and handling of non alphanumeric characters followed by tokenization. Tokens of size (number of characters) less than 3 were also removed due to a very low probability of these becoming indicative features to the classification model. Labels corresponding to sentiment classes of each segment in the dataset are mapped to the generic taxonomy classes (as discussed in Section 4.1), and the corresponding one-hot encodings are generated.

Text vectorization Using the pre-trained word2vec model (discussed in Section 4.2), each token is converted to a 200-dimensional vector. They are stacked together and padded to form a 2-D matrix of desired size (150 x 200). The number 150 denotes the maximum number of tokens in any preprocessed forum posts belonging to the training set.

4.4 UMLS Concept Retrieval

Identification of medical information like diseases, symptoms and treatments mentioned in a forum post is essential for the top-n similar post retrieval (Section 3.2) and treatment suggestion (Section 3.3) phases of the proposed framework. The Unified Medical Language System[7] (UMLS) is a compendium of many controlled vocabularies in the biomedical sciences (created in 1986). Therefore UMLS concept identifiers, related to the above mentioned medical information, were retrieved using Apache cTAKES[8]. Concepts with semantic type 'Disorder or Disease' were added to the set of diseases-those with semantic types 'Sign or Symptom' were added to the set of symptoms and those with semantic types 'Medication' and 'Procedures' were added to the set of treatments.

4.5 Relevance judgement for similar post retrieval

Annotating pairs of forum posts with their similarity scores as per human judgment is necessary to evaluate how much the retrieved text is relevant. This corresponds to the evaluation of the proposed custom similarity metric (E.q. 5). Since annotating every pairs of posts is a cumbersome task, 20% of the total posts in the dataset were randomly selected, maintaining equal class distribution for the annotation purpose. For each such post, top 5 similar posts are retrieved

[2] https://keras.io/
[3] http://scikit-learn.org/
[4] http://pandas.pydata.org/
[5] http://www.nltk.org/
[6] https://docs.python.org/3/library/pickle.html
[7] https://www.nlm.nih.gov/research/umls/
[8] ctakes.apache.org/

using the similarity metric. Annotators were asked to judge the similarity between each retrieved post and the original post on a Likert-type scale, from 1 to 5 (1 represents high dissimilarity while 5 represents high similarity between a pair of posts). Annotators were provided with the guidelines for relevance judgments on two questions–‘Is this post relevant to the original post in terms of medical information?’ and ‘Are the experiences and situations depicted in the pair of posts similar?’. A pair of posts is given high similarity rating if both of the conditions are true, and a low rating if neither is true. Three annotators having post-graduate educational levels performed the annotations.

We measure the inter-annotator agreement using Krippendorff’s alpha metric [9], and this was observed to be 0.78. Disagreements between the annotators can be explained on the basis of ambiguities, encountered during the labeling task. We provide few examples below:

1. There are cases where original writer of the blog assigned higher rating (denoting relevant), but the annotator disagreed on what constituted a ‘relevant’ post. This often corresponds to the posts giving general advice for illness. For example,’You can take xanax in case of high stress. It worked for me.’ Such advice may not be applicable to a certain specific situation.
2. Ambiguities are also observed for the cases where the authors of the posts are of similar age, sex and socio-economic backgrounds, but have different health issues (for example, one post depicted a male teenager with severe health anxiety, while the other post described a male teenager with social anxiety). For such cases, similarity ratings were varied.
3. Ratings also vary in cases where the symptoms match, but the cause and disorder differ. Annotators face problem in judging the posts which do not contain enough medical information. For example, headache can be a symptom for different diseases.

5 Experimental Results and Analysis

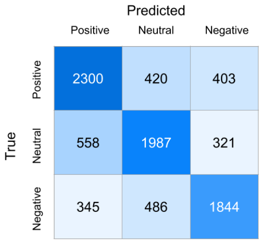
In this section, we report the evaluation results and present necessary analysis.

5.1 Sentiment Classification

The classification model (described in Section 3.1) is trained on a dataset of 8,864 unique instances obtained after preprocessing. We define a baseline model by implementing the CNN based system as proposed in [12] under the identical experimental conditions as that of our proposed architecture. We also develop

Model	Accuracy	Cohen-Kappa	Macro		
			Precision	Recall	F1-Score
Baseline [12]	0.63	0.443	0.661	0.643	0.652
LSTM	0.609	0.411	0.632	0.628	0.63
CNN-LSTM	0.6516	0.4559	0.6846	0.6604	0.6708
Proposed model	**0.6919**	**0.4966**	**0.7179**	**0.7002**	**0.7089**

Table 2. Evaluation results of 5-fold cross-validation for sentiment classification.



True \ Predicted	Positive	Neutral	Negative
Positive	2300	420	403
Neutral	558	1987	321
Negative	345	486	1844

Fig. 2. Confusion matrix of sentiment classification

a model based on an LSTM. To see the real impact of the third layer, we also show the performance of a CNN-LSTM based model. Batch size for training was set to 32. Results of 5-fold cross validation are shown in Table 2.

Evaluation shows that the proposed model performs better than the baseline system, and efficiently captures medical sentiment from the social media posts. Table 2 shows the accuracy, cohen-kappa, precision, recall and F1 score of the proposed model as 0.6919, 0.4966, 0.7179, 0.7002 and 0.7089, respectively. In comparison to the baseline model this is approximately a 9.13% improvement in terms of all the metrics. Posts usually consist of medical events and experiences. Therefore, capturing temporally related spatially close features is required for inferring the overall health status. The proposed CNN-LSTM-CNN network has been shown to be better at this task compared to the other models.

The high value of the Cohen-Kappa metric suggests that the proposed model indeed learns to classify posts into 3 sentiment classes rather than making any random guess. A closer look at the classification errors revealed that there are instances where CNN and LSTM predict incorrectly, but the proposed model correctly classifies. With the following example where baseline and LSTM both failed to correctly classify, but the proposed model succeeded: 'I had a doctors appointment today. He told I was recovering and should be more optimistic. I am still anxious and stressed most of the time'. Baseline model and LSTM classified it as positive (might be because of terms like 'recovering' and 'optimistic') while the proposed model classified it as negative. This shows that the proposed model can satisfactorily capture the contextual information, and leverage it effectively for the classification task.

To understand where our system fails we perform detailed error analysis- both quantitatively and qualitatively. We show quantitative analysis in terms of confusion matrix as shown in Figure 2. Close scrutiny of the predicted and actual values of test instances reveals that majority of misclassification occurs in cases where sentiment remains positive or negative throughout the post, and suddenly alters at the end of the post. For example: "I have been suffering from anxiety for a couple of years now. Doctors kept prescribing new and new medicines but I felt no change. Stress was unbearable. I lost my parents last year. The grief made me even worse. But I am currently feeling good after taking coQ10". We observe that the proposed model was confused in such cases. Moreover, users often share some personal content which does not help the medical domain significantly. Such noises also contribute to the misclassification.

Comparison with existing models: One of the recent works on medical sentiment analysis is reported in [12]. They have trained and evaluated a CNN based architecture separately for medical condition and medication segments. As discussed in the dataset and experiment section, we have merged both the datasets related to medical condition and medications into one for training and evaluation. Our definition of medical sentiment is, thus, more generic in nature, and direct comparison to the existing system is not very rational. None of the related works mentioned in the related works section addressed sentiment analysis for medical suggestion mining. The experimental setups, datasets and the sentiment classes used in all these works are also very different.

5.2 Top-N similar post retrieval

Evaluation of the retrieval task is done by comparing similarity scores assigned for a pair of forum posts by the system and by human annotator (as discussed in Section 4.5). Our focus is to determine the correlation between the similarity score assigned to the pairs of posts through human and the system judgments (rather than on actual similarity values). That is if a human feels that a post P is more relavant to post Q than post R, then the system also operates in the same way. Therefore, we use pearson correlation coefficient for the evaluation purpose. Statistical significance of the correlation (2-tailed p-value from *T-test* with null hypothesis that the correlation occured by chance) was found to be 0.00139, 0.0344 and 0.0186, respectively, for each sentiment class. Precision@5 is also calculated to evaluate the relevance of the retrieved forum posts. As annotation was done using top-5 retrieved posts (as discussed in Section 4.5), Recall@5 could not be calculated.

We design a baseline model using K-nearest neighbour algorithm that makes use of cosine similarity metric for capturing the textual similarity. We show the results in Table 3. From the evaluation results, it is evident that similarity scores assigned by the proposed system are more positively correlated with the human judgments than the baseline. Correlation can be considered statistically significant as the p-values corresponding to all the sentiment classes are less than 0.05. The better Precision@5 metric corresponds that a greater number of relevant posts are retrieved by the proposed approach in comparison to the baseline model.

We also calculate the Discounted Cumulative Gain (DCG) [7] of the similarity scores for both models from the human judgments. The idea behind DCG is that highly relevant documents appearing lower in the ranking should be penalized. A logarithmic reduction factor was applied to the human relevance judgment which was scaled from 0 to 4, and the DCG accumulated at a rank position 5 was calculated with the following formula:

$$DCG_5 = \sum_{i=1}^{5} \frac{rel_i}{\log_2(i+1)} \tag{7}$$

where rel_i is the relevance judgment of the post at position i. The NDCG could not be calculated, as annotation was done only using top-5 retrieved posts (as discussed in Section 4.5).

During error analysis, we observe few forum posts where users share their personal feelings, but due to the presence of less medically relevant contents, these are labeled as irrelevant by the system. However, these contain some relevant information that could be useful to the end users. For example, ’Hello everyone, Good morning to all. I know I had been away for a couple of days. I went outing with my family to get away from the stress I had been feeling lately. Strolled thorugh the park, played tennis with kids and visited cool places nearby. U know Family is the best therapy for every problem. Still feeling a little bit anxious lately. Suggest me something’ – The example blog contains proportionally more personal information than the medically relevant one. However, ’Feeling a little bit anxious lately’ is the medically relevant part of the post. Thus, filtering out such contents is required for better performance and would help the system to focus better on the relevant contents. There are two possible ways to tackle this problem. We would like to look into these in future.

1. Increasing the weight of medical information similarity (represented as *MISim* in Equation 5) while computing the overall similarity score.
2. Identifying and removing personal, medically irrelevant contents using (possibly) by either designing a sequence labeling model (classifying relevant vs. irrelevant) or by manually verifying the data or by finding certain phrases or snippets from the blog.

Sentiment	Pearson correlation		Precision@5		DCG$_5$	
	A	B	A	B	A	B
Positive	**0.3586**	0.2104	**0.6638**	0.5467	**6.0106**	2.1826
Neutral	**0.3297**	0.2748	**0.5932**	0.5734	**5.3541**	2.4353
Negative	**0.3345**	0.2334	**0.623**	0.5321	**4.7361**	2.4825

Table 3. Evaluation corresponding to the top-n similar posts retrieval. ’A’ and ’B’ denote the results corresponding to the proposed metric and K-nearest neighbor algorithm using text similarity metric, respectively.

5.3 Treatment suggestion

Evaluation of treatment suggestion is particularly challenging because it requires the annotators with high level of medical expertise. Moreover to the best of our knowledge there is no existing benchmark dataset for this evaluation. Hence, we are not able to provide any quantitative evaluation of the suggestion module. However, it is to be noted that our suggestion module is based on the soundness of sentiment classification module. Our evaluation presented in the earlier section shows that our sentiment classifier has acceptable output quality. The task of a good treatment suggestion system is to mine the best and relevant treatment suggestion for a candidate disease. As the function for computing the suggestion score (Eq. 6) involves computing the probability of positive sentiment, given a treatment T and disorder/disease D, it is always ensured that T is a candidate treatment for D, i.e. the treatment T produced positive results in context of D in at least one case. In other words, the probability term ensures that irrelevant treatments that did not give positive result in context of D would never appear as treatment suggestion for D. The efficiency of the suggestion module depends on the following three factors:

1. Apache cTAKES retrieved correct concepts in majority of the cases with only a few exceptions, which are mostly ambiguous in nature. For example, the word 'basis' can represent clinically-proven NAD+ Supplement or can be used as synonym of the word 'premise'.
2. If an irrelevant post is labeled as relevant by the system, then suggestions shouldn't contain treatments mentioned in that post. Thus, the similarity metric plays an important role in picking the right treatment for a given candidate disease.
3. Value of hyper-parameter τ (E.q. 6): As its value decreases, more number of candidate treatments are suggested by the system.

Performance of the module can be augmented and tailored by tweaking the above parameters depending on the practical application in hand.

6 Conclusion and Future Work

In this paper, we have established the usefulness of medical sentiment analysis for building a recommendation system that will assist building a patient assisted health-care system. A deep learning model has been presented for classifying the medical sentiment expressed in a forum post into conventional polarity-based classes. We have empirically shown that the proposed architecture can satisfactorily capture sentiment from the social media posts. We have also proposed a novel similarity metric for the retrieval of forum posts with similar medical experiences and sentiments. A novel treatment suggestion algorithm has been also proposed, that utilizes our similarity metric along with the patient-treatment satisfaction ratings. We have performed a very detailed analysis of our model.

In our work, we use the UMLS database due to its wide usage and acceptability as a standard database.We also point to other future work, such as annotating a dataset for treatment suggestions – which would increase the scope of machine learning, developing a sequence labeling model to remove personal irrelevant contents etc. Our work serves as an initial study in harnessing the huge amounts of open, useful information available on medical forums.

7 Acknowledgements

Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

1. Chee, B.W., Berlin, R., Schatz, B.: Predicting adverse drug events from personal health messages. AMIA Annu Symp Proc **2011**, 217–226 (2011)
2. Denecke, K.: Sentiment analysis from medical texts. In: Health Web Science: Social Media Data for Healthcare, chap. 10. Springer Publishing Company, Incorporated, 1st edn. (2015)
3. Denecke, K., Deng, Y.: Sentiment analysis in medical settings: New opportunities and challenges. Artificial Intelligence in Medicine **64**(1), 17 – 27 (2015). https://doi.org/https://doi.org/10.1016/j.artmed.2015.03.006, http://www.sciencedirect.com/science/article/pii/S0933365715000299
4. Eysenbach, G., Kohler, C.h.: What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet. AMIA Annu Symp Proc pp. 225–229 (2003)
5. Fox S, D.M.: Health Online. Washington, DC: Pew Internet & American Life; 2013. Jan 15, [2013-11-20] (2013), http://www.pewinternet.org/Reports/2013/Health-online/Summary-of-Findings.aspx
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, http://dx.doi.org/10.1162/neco.1997.9.8.1735
7. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (Oct 2002). https://doi.org/10.1145/582415.582418, http://doi.acm.org/10.1145/582415.582418
8. Kim, Y.: Convolutional neural networks for sentence classification. CoRR **abs/1408.5882** (2014)
9. Krippendorff, K.: Computing krippendorff's alpha-reliability. https://repository.upenn.edu/asc_papers/43 (2011)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: arXiv preprint arXiv:1301.3781,2013 (2013)
11. Palotti, J.R.M., Zuccon, G., Jimmy, Pecina, P., Lupu, M., Goeuriot, L., Kelly, L., Hanbury, A.: CLEF 2017 task overview: The IR task at the ehealth evaluation lab - evaluating retrieval methods for consumer health search. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017), http://ceur-ws.org/Vol-1866/invited_paper_16.pdf
12. Yadav, S., Ekbal, A., Saha, S., Bhattacharyya, P.: Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12, 2018 2018)
13. Yang, C.C., Yang, H., Jiang, L., Zhang, M.: Social media mining for drug safety signal detection. In: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing. pp. 33–40. SHB '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2389707.2389714, http://doi.acm.org/10.1145/2389707.2389714