Medical Forum Question Classification Using Deep Learning

Raksha Jalan, Manish Gupta[*], and Vasudeva Varma

International Institute of Information Technology-Hyderabad, India
jalan.raksha@research.iiit.ac.in, {manish.gupta,vv}@iiit.ac.in

Abstract. With the rapid increase in the number as well as quality of online medical forums, patients are increasingly using the Internet for health information and support. Online health forums play an important role in addressing consumers health information needs. However, given the large number of queries, and limited number of experts, a significant fraction of the questions remains unanswered. Automatic question classifiers can overcome this issue by directing questions to specific experts according to their topic preferences to get quick and better responses.

In this paper, we aim to classify health forum questions where classes of questions mainly focus on capturing user intentions. We strongly believe that a good estimate of user intentions will help direct their questions to the best responders. We propose a novel approach of combining medical domain based features with deep learning models for question classification task. To further improve performance of the data-hungry deep learning models, we resort to weak supervision strategies. We propose a new variant of the existing self-training method called "Self-Training with Lookups" for weak supervision. Our results demonstrate that combining features generated from biomedical entities along with other language representation features for deep learning networks can lead to substantial improvement in modeling user generated health content. Weak supervision further enhances the accuracy. The proposed model outperforms the state-of-the-art method on a benchmark dataset of 11000 questions with a margin of 3.13%.

Keywords: Medical Question Classification, Deep learning models, Weak Supervision

1 Introduction

With the increasing penetration of Internet to even the remote parts of the world, web based information access has become an integral part of meeting healthcare information needs. Unfortunately, search engines can still not answer a large number of healthcare queries effectively. Hence, recently, there has been a boom in the number of healthcare question-answering websites both specific to certain diseases, as well as general ones. To get better and quick responses to questions put up by an exponentially growing set of online information seekers, it is important to categorize questions and direct them to appropriate experts based on question types. As a first step in this direction we propose a novel approach for classifying questions posted on health forums into seven different categories each of which captures a unique user intent.

[*] The author is also a Principal Applied Scientist at Microsoft.

1. **Demographic** (DEMO): Questions targeted towards specific demographic subgroups characterized by age, gender, profession, ethnicity, etc.
2. **Disease** (DISE): Questions related to a specific disease.
3. **Treatment** (TRMT): Questions related to a specific treatment or procedure.
4. **Goal-oriented** (GOAL): Questions related to achieving a health goal, such as weight management, exercise regimen, etc.
5. **Pregnancy** (PREG): Questions related to pregnancy, difficulties with conception, mother and unborn child's health during pregnancy.
6. **Family support** (FMLY): Questions related to issues of a caregiver (rather than a patient), such as support of an ill child or spouse.
7. **Socializing** (SOCL): Questions related to socializing, including hobbies and recreational activities, rather than a specific health-related issue.

Most of the research reported in health domain has been performed on small datasets. Data annotation in general is expensive and time consuming. Specially in health domain, data annotation costs are significant since most of the annotations need to be done by medical experts. In such cases, weakly supervised techniques can lead to significant improvement in performance of various models. To the best of our knowledge, we are the first to introduce weak supervision for the question classification task in medical domain. In this paper we also propose a novel method for weak supervision: "self-training with lookups" which outperforms the typical self-training method [2].

Our proposed solution uses domain-specific knowledge by generating medical features along with word embeddings and word n-grams to train deep learning models in using the weakly supervised data. For building deep networks, we experiment with multiple sequence learning architectures including Hierarchical Bidirectional LSTMs [15] and fully connected neural networks. In order to capture the strength of multiple models, we train them separately and use weighted average ensemble method to generate final class predictions.

Overall, we make the following contributions in this paper. (1) We propose a novel approach to generate features in the medical domain for the question classification task. (2) We present a new variant of the self-training method called "self-training with lookups". (3) We demonstrate the use of weakly supervised data to train deep learning models which is very important in medical domain where we generally have limited annotated data. (4) Our best performing model provides an accuracy of 71.13%, which beats the state-of-the-art method with a margin of 3.13%. Data and code is available at `https://tinyurl.com/medCat18`.

The paper has been organized as follows. Related work is discussed in Section 2. We present details of our proposed methods in Section 3 followed by results and analysis in Section 4. Finally, we conclude with a brief summary in Section 5.

2 Related Work

Although there is a large amount of user generated content about healthcare on different social media sites, few studies have applied deep learning or artificial intelligence techniques for knowledge discovery on a large scale of data in this particular emerging area. In this section, we first discuss about recent advances in language modeling using deep learning, and then discuss about literature on question classification in healthcare.

2.1 Language Modeling using Deep Learning

Representation learning [7, 10] and deep learning models [6, 9, 15, 19] have been shown to be very effective for a large number of NLP and IR tasks [16]. Word embeddings [10] let you treat individual words as related units of meaning, rather than entirely distinct IDs. Benefiting from its recurrent structure, Recurrent Neural Networks (RNNs) [9] have been found to be very suitable to process variable length texts. But standard RNNs suffer from vanishing and exploding gradient problems. Long Short-Term Memory Networks (LSTMs) [6] deal with these problems by introducing memory cells and gates which allow for a better control over the gradient flow and enable better preservation of "long-range dependencies". Bidirectional LSTMs (BLSTMs) [17] utilize both the previous and future context by processing the sequence in both directions. Finally, hierarchical networks [15, 19] using two level BLSTMs have also been proposed. We use BLSTMs with attention to implement our baselines. Further, considering the hierarchical structure of questions, we build Hierarchical Attention BLSTMs for modeling questions. Our final model is based on ensemble techniques where we create an ensemble of multiple deep models generated using word embeddings and TF-IDF vectors.

2.2 Question Classification in the Health Domain

Liu et al. [8] classified medical domain questions according to whether they were asked by health-care professionals or consumers using statistical and category features to train SVMs. Roberts et al. [14] classified medical questions as Patient specific, general knowledge or research using lexical, syntactic and semantic features. Guo et al. [5] classified Chinese health-related questions into six categories (Condition Management, Healthy Lifestyle, Diagnosis, Health Provider Choice, Treatment, and Epidemiology) using lexical, grammatical, CMeSH concepts, keywords and statistical features. Mrabet et al. [12] identify topics from healthcare questions based on the contained entities. The question classification task discussed in this paper was first proposed at ICHI 2016[1]. The winner team [18] used TF-IDF features to train Logistic Regression, SVMs and Random Forests. They also trained a Convolutional Neural Network (CNN) for which they used pre-trained Google-news vectors. Final predictions were obtained by averaging the predictions generated by all four classifiers. They did not incorporate any domain based medical features into their models. Also, the small size of the dataset without usage of any weakly supervised or semi-supervised techniques leads to low accuracy. In this paper, we experiment with multiple feature sets, and multiple deep learning architectures along with weak supervision.

3 Proposed Methods

In this section we discuss our proposed methods for the question classification task. We start with a discussion of various pre-processing steps. Next, we mention various

[1] http://www.ieee-ichi.org/healthcare_data_analytics_callenge.html

standard sequence learning architectures which we adapt for our task. Further, we describe our medical features based approach, followed by our approach to perform weak supervision and ensemble learning.

3.1 Pre-processing

Data obtained from medical forums is usually noisy as it is user generated. Hence, we perform basic pre-processing and cleaning of the data before building classification models. We remove all hyper-links, special characters, punctuations and stopwords from the title and the body of questions. Then we perform case-folding and lemmatization. We concatenated the title and body of the question together for all questions before using it for our models. So the question dataset can be represented as $Q = \{q_1, q_2, q_3, ..., q_N\}$, where $q_i = t_i + b_i$, such that q_i corresponds to the i^{th} question and t_i and b_i are the title and the body of the question respectively. When training instances are limited, word embeddings generated using only training instances cannot capture the semantic meaning of words effectively. Hence, in our experiments we use pre-trained embeddings and fine-tune them during training to improve the performance of classification. We experiment with multiple pre-trained word embeddings to obtain the best representation for the questions. Among embeddings such as Wikipedia-Pubmed-and-PMC-w2v vectors [11], Google-News Vectors[2] and Global Vectors for Word Representation (GloVe) [13], we found GloVe to outperform others for our task. Hence, we report results using GloVe embeddings.

3.2 Standard Sequence Learning Architectures

Next, let us discuss multiple variants of the sequence learning framework which can be leveraged for our task.

3.2.1 Bidirectional LSTM (BLSTM) Bidirectional LSTMs train two LSTM models together on the input sequence. The first is trained on the input sequence itself, and the other on a reversed copy of the input sequence. This provides both forward as well as backward context to the network and results in faster and more holistic learning on the problem. Each question is represented using a matrix with each row corresponding to a GloVe vector for a word in the question. This matrix is passed to BLSTM model, which generates the encoded representation of question which is then passed as input to softmax layer for classification. We use Bidirectional LSTM with 128 dimensions in each direction.

3.2.2 Bidirectional LSTMs with Attention Networks (BLSTM-A) "Attention Mechanism" has been proposed recently, for effective modeling of long-term dependencies. Attention mechanisms allow for a more direct dependence between the state of the model at different points in time. Questions are represented using a matrix same as that for BLSTM network which is then passed to a BLSTM to generate encoded representation of questions. Attention layer is built on top of the Bidirectional-LSTM layer (128) followed by softmax layer at the top for classification.

[2] https://code.google.com/archive/p/word2vec/

3.2.3 Hierarchical Networks So far, we represented a question as a sequence of words and encoded it using a matrix. Better representations for questions of larger length can be obtained by incorporating knowledge of their structure in the model's architecture. When questions contain multiple sentences, each question can be considered as a sequence of sentences and each sentence as sequence of words. In this style, questions are represented using a 3D tensor where the additional third dimension represents sentences in the question. Each of these sentences are then represented using the typical 2D matrix generated from sequence of words from the sentence and their corresponding GloVe vectors.
Hierarchical Bidirectional-LSTM (H-BLSTM): For this model, each question is encoded in a hierarchical fashion. Sentences within question are first encoded using sentence encoders (using Bidirectional-LSTMs) and then these encoded sentences are passed to a document (question in our case) encoder (again using another Bidirectional-LSTM), which encodes the question, and generates a final encoding for the question. This encoded representation is then passed to a softmax layer for classification. We used same number of LSTM dimensions (128) for both sentence encoder and document encoder.
Hierarchical Attention Networks (HAN): Not all sentences in a question are equally informative for representing a question and determining the informative sentences involves modeling the interactions of the words, not just their presence in isolation. We implemented a hierarchical network with two levels of attention mechanisms applied at the word and sentence-level, enabling it to attend differentially to more and less important content when generating the question encoding. Finally encoded representation of question is passed to a softmax layer for classification.

3.2.4 Term frequency-Inverse Document Frequency based Deep Network (TFIDF-DN) Term frequency-Inverse Document Frequency (TFIDF) [3] is one of the popular feature representation methods for text which can efficiently capture statistical properties of words. We represented questions with their TF-IDF vectors and trained neural network consisting a fully connected layer followed by a dropout layer with a softmax layer at the top.

3.3 Deep Model Based on Medical Features

In order to incorporate domain knowledge and important medical signals, we use MetaMap [1] which is primarily designed to extract medical entities from biomedical documents. It gives us the advantage of capturing critical medical features based on multi-word combinations which are difficult to capture using single word based methods proposed in the previous sub-section. For example: “Heart Attack”, “Chronic Fatigue Syndrome”, “Chest pain”, etc. By relating such entities with the class labels, we generate “association strength” based medical features for each question as discussed in the following.

MetaMap also provides semantic mappings for medical entities. It has a total 133 types of semantic mappings. However we only consider those medical entities for feature generation that belong to at least one of the following 16 semantic mappings, since these are relevant for our task. This is important since including entities from other semantic mappings could lead to noisy entity linking and hence poor feature generation.

Algorithm 1 Generating Features based on SoA Scores of Medical Entities

```
Input  Entities list $l_i$ for each $q_i \in Q$, threshold $\tau$, Entity level SoA scores $S_e^{M \times K}$
Output  Question level SoA scores $S_q^{N \times K}$
Initialization: $S_q = [0]^{N \times K}$
for $i = 0$ to $N$ do
    for all $entity \in l_i$ do
        $(max_1,\ max_2) \leftarrow$ (Maximum, Second Maximum) value in $S_e[entity]$
        if $\frac{max_1}{max_2} > \tau$ then
            $index \leftarrow argmax(S_e[entity])$
            $S_q[i][index] = S_q[i][index] + 1$
Return $S_q^{N \times K}$
```

Types of semantic mappings we used for feature generation are: Antibiotics, Clinical Drugs, Diagnostic Procedures, Indicator-Reagent, Diagnostic Aid, Therapeutic Procedure, Drug Delivery Device, Anatomical Abnormality, Disease and Syndrome, Sign or Symptom, Family Group, Body System, Biological Region or Location, Biological Function, Body Parts, Body Space, and Age Group.

For each question $q_i \in Q$, we obtain a list l_i of medical entities belonging to any of the above semantic mapping types. Next, for each entity, we compute their Strength of Association (SoA) scores with every class label. Let K be the number of classes (in our case, K=7), and let us denote the class labels by $\{c_i\}_{i=1}^{K}$. We define the Strength of Association (SoA) between an entity e and a class label c_i as follows.

$$SoA(e,c) = \log_2 \frac{P(e|c)}{P(e|\neg c)} = \log_2 \frac{freq(e,c) \times freq(\neg c)}{freq(e,\neg c) \times freq(c)} \quad (1)$$

where $freq(e,c)$ is the number of questions with label c which contain the entity e. $freq(c)$ is the number of questions in class c. $freq(e,\neg c)$ is the number of times e occurs in questions in classes other than c. $freq(\neg c)$ is the number of questions with label other than c. The intuition is that the entity e is associated closely with class c if it occurs in questions labeled c much more number of times compared to occurrences in questions of another class.

Though association of individual words with the class labels is captured by TF-IDF, SoA captures the association between domain-specific multi-word medical entities and class labels. If an entity has a stronger tendency to occur in a question with a particular label than in questions with other labels, then that (entity, label) pair will have an SoA score greater than zero. Let M be the total number of unique medical entities in the dataset. Since there are K (=7) class labels, we obtain a SoA matrix $S_e^{M \times K}$. Algorithm 1 illustrates the method we use to process these entity-level SoA scores to come up question-level SoA scores $S_q^{N \times K}$ where N is the number of questions in the dataset. For each question q_i, we first retrieve the list of medical entities l_i. Then for each entity in the list we find the maximum (max_1) and the second maximum (max_2) SoA scores for the entity. We intend to capture only strong associations between entities and labels. Hence, only if the ratio of max_1 to max_2 is greater than a defined threshold τ, we let the entity e contribute to the SoA score for the question q_i. The SoA scores in S_q are the medical features fed to a fully connected layer followed by a combination of a dropout layer and a softmax layer.

3.4 Models using Weak Supervision

First, we crawl $\sim$100K health questions from “medhelp.org”[3] to gather a large dataset of unlabeled questions. After pre-processing the crawled questions (as discussed in Section 3.1), we use self-training to generate weakly supervised labels. Self-training is a popular method for weak supervision. It is an iterative method where every iteration k contains two steps: (1) based on the current labeled dataset L_{k-1}, a classifier C_k is trained, and (2) the classifier C_k is used to predict labels for instances in the current unlabeled data U_{k-1} leading to classification of more instances to generate the new U_k and L_k.

Two techniques are popularly used for robust self-training:

- Throttling Principle: An unlabeled instance $x \in U_{k-1}$ is assigned to a class c_i (and hence to L_k) in the second step only if the classifier C_k predicts the class c_i for x with a probability greater than a threshold T.
- Balancing Principle: In order to avoid class imbalance, equal number of instances of all predicted classes are added to the labeled data for the next iteration.

One disadvantage of the traditional self-training method is that a particular unlabeled instance might be predicted to belong to a class with high confidence incorrectly just because the parameters of the model have not yet stabilized and are still being learned themselves. This could lead to propagation of errors over iterations due to corruption of training data. This is particularly critical in the first few iterations where the model has been learned using a very small labeled dataset and is therefore “weak”.

In this paper, we propose a novel variant of the self-training method called “self-training with lookups” to handle this problem. Besides following the principles of throttling and balancing, it makes use of lookup lists which significantly decreases incorrect label assignments to unlabeled data compared to traditional self-training. We maintain one lookup list for each class. The proposed self-training with lookups method is illustrated in Algorithm 2. In our algorithm, if the classifier predicts label c for instance $u \in U$ with probability greater than the throttling threshold T for the first time, the instance is not directly added to the labeled set. Instead, it is added to a lookup list for the class c. However, if the instance u is labeled with the class c again in subsequent iterations (and so the lookup list for the class c already contains u), it is added to the labeled set with label as c.

On detailed analysis, we find that the proposed method provides significant reduction in noisy labels compared to the traditional self-training method, at the cost of a few more computations. In order to compare the performance of the proposed method with the traditional self-training method, we use exactly the same parameters for both. Out of 100000 unlabeled questions, our approach assigns labels to 34664 questions while the traditional self-training algorithm generates labels for 53123 samples after the same number of iterations (10). Further, we compare the quality of weakly supervised question labels generated from both the algorithms by using the generated labeled data to learn initial parameters of various deep learning models followed by fine tuning with the original labeled training data.

[3] `http://www.medhelp.org/`

Algorithm 2 Self-training With Lookups

```
Input  Labeled Dataset $L = \{\langle x_i, y_i \rangle\}_{i=1}^{M}$, Unlabeled Dataset $U = \{x_i\}_{i=1}^{N}$, Throttling threshold $T$
Output  Classifier $C$, Augmented Labeled Dataset $L$.
Initialization: Lookup Lists $l[c] = \phi \;\; \forall c \in \{c_1, \ldots, c_K\}$.
while stopping criteria is not met do
    $C \leftarrow Train\_Model(L)$
    Candidates Cand$[c] = \phi \;\; \forall c \in \{c_1, \ldots, c_K\}$.
    $Pred \leftarrow Predict(U, C)$ where $Pred[u][c]$ is the probability of classifying instance $u \in U$ to class $c$.
    for each $u \in U$ do
        $r \leftarrow argmax(Pred[u])$
        if $Pred[u][r] > T$ then                                  ▷ Throttling
            if $u \in l[c_r]$ then
                Cand$[c_r] \leftarrow$ Cand$[c_r] \cup \{u\}$                     ▷ Lookups
            else
                $l[c_r] \leftarrow l[c_r] \cup \{u\}$
    $size \leftarrow \min(|$Cand$[c_1]|, \ldots, |$Cand$[c_K]|)$                ▷ Balancing
    for each $c \in \{c_1, \ldots, c_K\}$ do
        for $i \leftarrow 1$ to $size$ do
            $L \leftarrow L \cup \{\langle$Cand$[c][i], c\rangle\}$
            $U \leftarrow U - \{$Cand$[c][i]\}$
Return C, L.
```

While the size of the labeled data generated by our method is expectedly small, the quality of labels using our method leads to better classification accuracy. Table 1 shows the comparison results. Accuracy values in each cell are the average values obtained after repeating the experiments five times. For all the models, experiments with labels generated using our approach performed better than the ones with labels generated using the traditional self-training method. Self-training with lookups outperforms traditional self-training method by ~1% on average. Hence, for all further experiments, we use weakly supervised labels generated using the "self-training with lookups" method.

Model Name	Without Lookups	With Lookups
BLSTM	64.47 %	65.01%
BLSTM-A	64.75%	65.67%
H-BLSTM	68.09%	68.35%
HAN	67.10%	67.98%
TFIDF-DN	67.05%	68.96%

Table 1: Comparison of Self-training with Lookups and Traditional Self-training

3.5 Overall Model using Ensemble Learning

In recent years, ensemble based methods have been found to be successful across multiple machine learning tasks such as classification, clustering, anomaly detection, etc. For classification, it has been well studied that combining models which individually use different features can result in accuracy gains since different feature spaces result into uncorrelated errors. In our case, we combine the benefits of memorization (Deep Networks), generalization (TF-IDF based models) and domain knowledge (SoA based model) to build an ensemble model to get the final class prediction. We experiment with various ways of combining these individual methods including cascading, averaging, and weighted averaging. Due to lack of space we will list the results obtained using the top performing models only. Figure 1 shows our overall model architecture.

In supervised learning only labeled training instances are used to train the model. While in weakly supervised methods, the model is first trained using weakly supervised data and then labeled training data is used to fine tune the model. In Figure 1, W_{sk} represents the 2-Dimensional embedding matrix of the k^{th} sentence in a question. The embedding matrix W_{sk} is fed to a Bidirectional LSTM B_k which creates an encoding for the corresponding input sentence. Each block B_k follows the same internal structure as shown in the figure where multiple LSTM cells (L) are connected in forward and reverse fashion. Further, the encoded sentences for the question generated by the first level of BLSTM are input to another BLSTM (B_Q) which encodes the complete question. Then this question encoding is passed to the softmax layer for classification. For the SoA based deep network and the TF-IDF based deep network, as mentioned earlier, the representations are fed to a fully connected layer followed by a combination of a dropout layer and a softmax layer at the top. The class predictions P_{SOA}, P_{HBLSTM}, P_{TF-IDF} obtained from the three models are combined using weighted averaging to obtain the final prediction P_{FINAL}. For ensemble based models, we report accuracies obtained using the weighted average scheme [4]. We fine tuned the parameters (W_{we}, W_{tfidf}, W_{soa}) which represent the weights given to predictions generated by the models, using the original labeled dataset. Note that W_{we} denotes the weight to the word embedding based HBLSTM model, W_{tfidf} denotes the weight to the TF-IDF based model and W_{soa} denotes the weight to the SoA based model.

Results obtained using this ensemble model outperform the baselines in both supervised as well as weak supervised settings. We discuss the performance details in Section 4.

4 Results and Analysis

The dataset provided as part of the Healthcare Data Analytics Challenge at ICHI 2016[4] contains real questions posted on a health discussion forum. The training data has 8000 questions and the test data contains 3000 questions each with the question title, question text, and a category. The categories are: (1) Demographic, (2) Disease, (3) Treatment, (4) Goal-oriented, (5) Pregnancy, (6) Family support, and (7) Socializing. The dataset is balanced with almost same number of instances per class. The performance results of various models in the supervised and the weakly supervised settings are shown in Table 2 and Table 3 respectively. All accuracies are computed on the test set of size 3000, provided in ICHI data analytics shared task.

4.1 Results of Supervised Models

Table 2 shows the results obtained using various supervised models. We observe that in the pure supervised environment, models built on word embedding based features

[4] http://www.ieee-ichi.org/healthcare_data_analytics_callenge.html

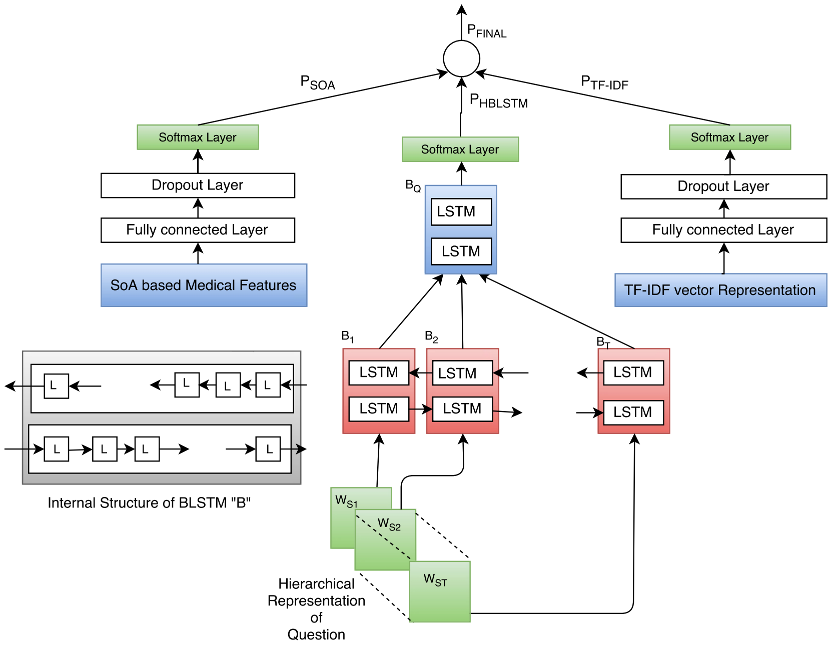

Fig. 1: Architecture of the proposed Deep Learning based Question Classifier

or TF-IDF based features could not beat the accuracy of the ICHI 2016 winners model which is the current state-of-the-art for this task. However, the ensemble based method with weighted combination of predictions from the models trained on SoA based medical features along with the predictions from word embedding and TFIDF based models, leads to an accuracy which is significantly better.

For the case, when we used hierarchical-BLSTMs for word embeddings, we found the best weights as $W_{we} = 0.73$, $W_{tfidf} = 1.15$, $W_{SoA} = 0.66$. We observed that the SoA based model improved the overall performance of the model by 3.02%. For SoA based model, we chose the threshold τ=1.35 after tuning it between 0.5 and 1.5. When we changed the word embedding model to HANs, we found the following weights for the ensemble model $W_{we} = 0.70$, $W_{tfidf} = 1.15$, $W_{SoA} = 0.66$. In this case, we observed that the SoA based model improved the overall performance of this model by 2.47%.

4.2 Results of Weakly Supervised Models

Table 3 shows the results obtained using various weakly supervised models. Distance supervision using weak supervision based methods has been shown to improve overall performance of deep learning methods for a large variety of tasks. We obtain large amounts of labeled data using weak supervision on data obtained from “medhelp.org”. Pre-training of the network parameters using such weakly supervised data improved the individual models' accuracies as shown in Table 3. This also resulted in improved accuracies for the ensemble based models.

In our study, we achieved the highest accuracy of 71.13% using the ensemble model (in the weakly supervised setting) which combines Hierarchical-BLSTM, TF-IDF based classifier, and the SoA based classifier. This also outperformed all other models in the supervised setting. It beats the current baseline model by a margin of 3.13%.

Table 2: Accuracy Comparison of Various Supervised Models

Models	Accuracy
ICHI 2016 Challenge Winners [18]	**68.00%**
BLSTM	62.70%
BLSTM-A	62.53%
H-BLSTM	64.03%
HAN	63.93%
SoA-DN	59.76%
TFIDF-DN	65.10%
HAN + TFIDF-DN	66.37%
H-BLSTM + TFIDF-DN	66.74%
HAN + TFIDF-DN + SoA-DN	**68.84%**
H-BLSTM + TFIDF-DN + SoA-DN	**69.76%**

Table 3: Accuracy Comparison of Various Weakly Supervised Models

Models	Accuracy
H-BLSTM	**68.35%**
HAN	67.98%
TFIDF-DN	**68.96%**
HAN + TFIDF-DN + SoA-DN	**70.37%**
H-BLSTM + TFIDF-DN + SoA-DN	**71.13%**

4.3 Error Analysis

We observed that for instances which contain strong features indicative of multiple classes, sometimes the classifier fails to predict the most dominant class. For example, consider the question "Last night I got high fever. My dad consulted his doctor friend....". The most dominant category for this question is the "Disease" category but because of the terms like "My dad", "consulted", "his doctor", it got categorized as "Family". Another kind of questions where the classifier fails is when the forum question is about particular disease symptoms, and in between the question mentions possible treatment plans for that specific disease. In such cases sometimes the classifier fails to predict the correct label. Perhaps modeling the problem in a multi-label multi-class classification setting will address such problems.

5 Conclusions

In this paper, we discussed the problem of medical forum question classification. We found that using domain knowledge based features along with word embeddings provides better accuracy compared to just using word embeddings in the deep learning supervised setting. Biomedical entities can lead to substantial improvement in modeling the user generated health content. In addition to the proposed usage of medical features, we experimented with various methods to generate weakly supervised labels, and presented our new approach "self-training with lookups". Our experiments demonstrate the effectiveness of the ensemble of the three models (word embedding based sequence learning models, model with SoA based medical domain features, and TF-IDF based statistical model). As part of future work we plan to pursue two directions: (1) Explore the lookups idea further by trying combinations like varying the number of iterations in which the classifier assigns a label to an instance (currently set as 2), and varying throttling threshold over iterations. (2) Model other forms of context related to the question, for example, user who asked the question, time when the question was asked, etc.

References

1. Aronson, A.R.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In: AMIA. p. 17 (2001)
2. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised Learning. IEEE Trans. on Neural Networks 20 (2009)
3. Christopher, D.M., Prabhakar, R., Hinrich, S.: Introduction to information retrieval. An Introduction To Information Retrieval 151, 177 (2008)
4. Dietterich, T.G., et al.: Ensemble Methods in Machine Learning. Multiple Classifier Systems 1857, 1–15 (2000)
5. Guo, H., Na, X., Hou, L., Li, J.: Classifying Chinese Questions related to Health Care posted by Consumers via the Internet. Journal of Medical Internet Research 19(6) (2017)
6. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural computation 9 (1997)
7. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. In: ICML (2014)
8. Liu, F., Antieau, L.D., Yu, H.: Toward Automated Consumer Question Answering: Automatically Separating Consumer Questions from Professional Questions in the Healthcare Domain. Journal of Bio. Info. 44 (2011)
9. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent Neural Network based Language Model. In: Interspeech (2010)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: NIPS (2013)
11. Moen, S., Ananiadou, T.S.S.: Distributional Semantics Resources for Biomedical Text Processing (2013)
12. Mrabet, Y., Kilicoglu, H., Roberts, K., Demner-Fushman, D.: Combining Open-domain and Biomedical Knowledge for Topic Recognition in Consumer Health Questions. In: AMIA. vol. 2016, p. 914 (2016)
13. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: EMNLP. pp. 1532–1543 (2014)
14. Roberts, K., Rodriguez, L., Shooshan, S.E., Demner-Fushman, D.: Resource Classification for Medical Questions. In: AMIA (2016)
15. Ruder, S., Ghaffari, P., Breslin, J.G.: A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. arXiv preprint arXiv:1609.02745 (2016)
16. Socher, R., Bengio, Y., Manning, C.D.: Deep learning for nlp (without magic). In: Tutorial Abstracts of ACL 2012. pp. 5–5. Association for Computational Linguistics (2012)
17. Tan, M., Santos, C.d., Xiang, B., Zhou, B.: LSTM-based Deep Learning Models for Non-Factoid Answer Selection. arXiv preprint arXiv:1511.04108 (2015)
18. Verma, J., Kwon, B.C., Cheng, Y., Ghosh, S., Ng, K.: Classification of Healthcare Forum Messages. In: ICHI (2016)
19. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J., Hovy, E.H.: Hierarchical Attention Networks for Document Classification. In: HLT-NAACL. pp. 1480–1489 (2016)