IRE MAJOR PROJECT

Group No 27

Members:

Sejal Naidu (20162104)
Ramamurthi Kumar (20162029)
Abhishek Garg (201401020)
Diplav Srivastava (201430116)

Mentor - Raksha Jalan

Classification of Health Forum Messages using Deep Learning

PROBLEM STATEMENT

Given set of queries from health forum, task is to classify them into one of the 7 categories. The different categories are :

1) Demographic (DEMO)
2) Disease (DISE)
3) Treatment (TRMT)
4) Goal-oriented (GOAL)
5) Pregnancy (PREG)
6) Family support (FMLY)
7) Socializing (SOCL)

Applications

- This model can be used by pharmaceutics and health related websites to classify their blogs and public posts to allow consumers to easily browse through them.
- The users can easily search for posts based on category. Eg. Pregnancy, family support, demographic etc.
- Information seekers post their questions on forums, and other members of communities provide their suggestions or answers. Tagging the posts with relevant tasks helps draw attention from the members with relevant expertise.

Challenges

- Data cleaning - Data has unnecessary information like hyperlinks,non-space separated words,mixed case words
- Building feature vector - how to make feature vector containing semantic information related to medical terms
- Identifying different models according to the given problem statement

Dataset

The data used for this work is provided as a part of Healthcare Data Analytics Challenge at ICHI 2016, which contain real messages posted on a health discussion forum. Two different data files were provided in tab separated format for training and testing, respectively. The training data has 8000 messages each with the title text, contents text, and a category.

Preprocessing

- We performed a high-level preprocessing and cleaning of data before building our classification model. We started with removing all the stop-words, hyperlinks and special characters from the text.
- Then we built features for each question entry by extracting unigram and bigram tokens.
- We transformed the raw tokens through term frequency-inverse document frequency (tf-idf) transformation.
- These tf-idf vectors were used by most of our approaches.

Word2vec

Word2vec is a two-layer neural net that processes text. Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus.
While Word2vec is not a deep neural network, it turns text into a numerical form that deep nets can understand.
We used word2vec to extract features from given data and preprocess the dataset.

Doc2vec

Doc2Vec is used to associate arbitrary documents with labels, so labels are required. Doc2vec is an extension of word2vec that learns to correlate labels and words, rather than words with other words. The first step is coming up with a vector that represents the “meaning” of a document, which can then be used as input to a supervised machine learning algorithm to associate documents with labels.

We used both word2vec and doc2vec to improve the accuracy of our model.

Model : DNN

We first tried the standard feed-forward neural network. Our input is the word vector representing each of the message x . After forward propagating the input through the deep network, the Output y would be a 7-dimensional one hot encoding corresponding to each of the class.

Model : CNN

We also explored a convolutional neural network trained in a one-vs-all fashion. These models can better incorporate long range context in the text to be classified, so they can outperform the baseline competitors. Here, we used a network with convolutional filter window sizes of 4, 5, and 6 for layer 1,2, and 3 respectively. The number of hidden units was chosen to be 200. The Rectified linear unit (RELU) non-linear activation function was used. The outputs of these were max pooled over time and fed into a logistic output layer. The parameters of the network were learned in mini-batches of size 50 using adadelta a method for gradient descent which adapts dynamically and has minimal computational expense, with the value of decay parameter 0.95. To guard against overfitting, we also employed dropout and l2 regularization with strength 3.

Model : LSTM

To explore more advanced model, we adopted the LSTM model introduced by Hochreiter & Schmidhuber. In this model, we used a trained word2vec model on a global word-word co- occurrence matrix(all words in all reviews. And then we feed our representation vectors into LSTM model.

Result

Models	Count Vectorizer	tf-IDf	Word2Vec	Doc2Vec
DNN	54	60	33	47
CNN	-	-	39	50
LSTM	-	57	14	15

Conclusion

The goal of this project is to explore the architecture of neural network models that best suits classification of health forum messages and automatically tag the messages with their most probable category. The results show descent accuracy of 62%. One model tends to surpass another by having more parameters at the price of more training time. The biggest bottleneck of the model is to handle the medical terms not in dictionary. To further improve the performance, we can simply increase the memory for large vocabulary or have a hybrid word character model that switches to character based classification when faced with unknown words.

References

- Classification of Healthcare Forum Messages - Janu Verma, Bum Chul Kwon, Yu Cheng, Soumya Ghosh, Kenney Ng
- Deep Learning for Query Semantic Domains Classification - Ting Fang