Understanding the Sentiment of Clinical Notes

Jiaming Zeng
jiaming@stanford.edu
Department of Management Science and Engineering
Mentor: Suvadip

Abstract

Sentiment analysis has been a huge field of NLP. However, there isn't much sentiment analysis done for clinical notes. For my project, I performed sentiment analysis on clinical notes by building off the existing sentiment analysis framework for movie reviews. By mixing self-annotated clinical notes data with movie review data, I was able to train the new sentiment analysis algorithm.

1 Approach

Sentiment analysis is one of the primary fields of natural language processing. People have used various sentiment analysis algorithms to classify the sentiment in reviews and opinions. However, not much work has been done in the field of sentiment analysis for medical notes. Doctors writes copious amount of notes and this information is recorded in electronic medical health records (EMR). My project attempts to understand the sentiment of these notes using word2vec models and simple deep learning training. The project will be helpful in understanding the progression of a patient through the clinical notes and also explores how well the sentiment analysis algorithms developed for other reviews translate to the clinical setting.

The major challenges to sentiment analysis for clinical notes are 1) lack of correct annotation for sentiment analysis 2) the complexity and mixture of content available in clinical notes. Many times, clinical notes are written in an objective manner and are meant more to describe events than to convey opinion. Moreover, clinical notes does not only record opinions but also procedures, directions, and general information for the patients. In summary, the notes are extremely messy and it would take a doctor a long time to go through and annotate them. Sentiment analysis could be very interesting for clinical notes because we can see if we can track the progress of the disease through the doctor's sentiments.

The existing literature on sentiment for clinical settings and unsupervised sentiment analysis is limited. For sentiment analysis on clinical settings, there are existing work done in [1] and [2]. In [1], they had doctors hand-annotate the notes and performed analysis. [1] found that medical sentiment analysis requires a domain-specific sentiment source and many implicit sentiment needs must be considered. In [2], they used word2vec to perform unsupervised sentiment analysis based on clinical notes. However, the point was more comparing sentiments across different groups of notes instead of analyzing the sentiment of each individual notes. In the space of unsupervised sentiment analysis, most of the existing algorithms rely on clustering algorithms, such as [3], [4], [5].

For the project, instead of using the approaches already experimented with above, I plan to tackle the problem by building on the existing sentiment analysis framework for movie reviews. Sentiment analysis for movie reviews has been a well-studied problem and a standard word2vec model achieves relatively stable and high training accuracy. I'll create some hand-annotated notes from the clinical narratives and combine that with the existing IMDB movie review database. Then I'll train the models as normal and examine the performance of the sentiment analysis on the new dataset mixed with clinical data.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2 Experiments

2.1 Data

We used the Stanford Cancer Institute Research Database (SCIRDB) for the clinical notes. From the database, I pulled 53 patients with localized cancers with a total of 8,376 notes. The notes are then segmented into rough sentences and phrases based on the spaces. Due to the fact that clinical notes includes much introductory information that do not qualify as sentences and also signatures on the bottom. I cut out the first 4 and last 2 sentences. Moreover, to tackle the problem that many of the sentences or phrases are descriptions with no opinion, I only sentences with more than 50 characters for analysis. With that, I was able to cut out some short lab result readings and nondescript information. Finally, because there is no annotation available, I hand-annotated 299 sentences selected based on the above method.

For the movie review database, I used the IMDB database used in [6]. I extracted and processed the entire IMDB dataset based on the [7]. In the end, I was able to extract 50,000 sentences from the movie review. The two dataset was combined to form a new mixed dataset of 50,299 sentences. For training and testing, I used a split of 80% training and 20% testing.

2.2 Evaluation Method

I will evaluate the methods by 1) examining the training and testing accuracy from the model and 2) I will examine the performance of the model on new clinical note sentences that were not used for training. Hence, I will evaluate the success of sentiment analysis by looking at the results myself and seeing whether the sentences were classified correctly.

2.3 Experimental Details

For the experiment, I followed the procedure and code presented in [8]. Instead of performing everything with just movie review data. I used my mixed dataset of clinical and movie notes. The network structure I used was taken from [8], as seen in Figure 1.

```
Summary of the model...
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 2678, 100)         13415700
_________________________________________________________________
gru_3 (GRU)                  (None, 32)                12768
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 33
=================================================================
Total params: 13,428,501
Trainable params: 12,801
Non-trainable params: 13,415,700
_________________________________________________________________
```

Figure 1: Architecture of neural network.

To start, I pre-trained word2vec embeddings on only the movie review data. The movie review data was better processed and I wanted to see how the word embeddings trained on the movie data translated to the clinical notes. The pre-trained word embeddings were then tokenized and passed into the embedding layer of the neural network. This reduces the number of trainable parameters in the neural network and should train faster.

Then, the mixed dataset is split into training (80%) and testing (20%). I then padded all the sentences in the mixed dataset and trained for two different settings: 1). 25 epochs with 128 batch size, and 2). 100 epochs with 64 batch size. The trained model is then evaluated on 10 sentences chosen from the set of untrained clinical notes. The results are examined in the next section.

Again, the code and experiment performed here is based heavily on the example presented in [8].

2.4 Results

I'll analyze the results from two aspects: 1) training and testing accuracy from the model. 2) the performance on the 10 sentences chosen from the untrained clinical notes.

For the training and accuracy, the model did not perform very well for either the 25 epochs setting or the 100 epochs setting. The results on loss and accuracy can be seen in the Table 1. The results from the original example in [8] are also presented for comparison purposes.

Setting	Loss	Training Accuracy	Testing Accuracy
mixed data (25 epochs)	0.66	0.59	0.5
mixed data (100 epochs)	0.65	0.62	0.5
movie data (5 epochs)	0.32	0.86	0.88

Table 1: Results from the mixed dataset and original movie data for sentiment analysis.

As we can see, the model with the clinical data mixed in completely broke down the sentiment analysis model that was working very well on the movie data. The training accuracy only increased by 0.03 even though I quadrupled the training time. Moreover, there is absolutely no effect on the testing accuracy and the model does no better than a random guess. This was fairly surprising to me as the majority of the data was still movie review data. In fact, there were only 299 clinical notes sentences out of the total 50,299 sentences. Because of my lack of clinical background, the clinical notes data may have been badly processed and labeled. However, it was surprising to me that with just the addition of 0.5% of "adversarial" data, I was able to completely break down a previously well-performing sentiment analysis model.

To continue with my planned evaluation methods, I examined the sentiment score for 10 untrained sentences from the clinical notes. The results are detailed in Tables 2 and 3. As seen from the training and testing accuracy results, the model has no idea how to classify each of the clinical sentences. Examining the sentences themselves, it should also be noted that the vocabulary is extremely foreign to the movie review trained word embeddings and the language does not really lend itself to tell us whether the sentence is "positive" or "negative".

Hence, it seems like the results from the mixed dataset does not make sense and the addition of 0.5% of "adversarial" data has completely broken down a well-established sentiment analysis framework.

Text	Score
he initially presented with gross hematuria in about november and had a ct scan that showed no hydronephrosis but tumor overlying the left posterior wall	0.628
turbt was read as showing high grade tumor with widespread invasion of the lamina propria but no definite detrusor muscle in the specimen	0.519
it sounds like dr nickas was suspicious of muscle invasion	0.515
about 2 weeks after the turbt he had clot retention and was admitted to john muir hospital for a few days and was cared for by dr hopkins requiring catheter placement and irrigation of the bladder	0.561
repeat turbt by dr skinner on 4 10 17 showed a left anterior wall bladder tumor	0.532
resection of tumors of the left anterior wall left lateral wall and tumor base were performed	0.394
pathology showed small cell carcinoma 95 urothelial carcinoma with glandular differentiation 5 invasive of lamina propria and muscularis propria	0.549
staging pet ct on 4 21 17 showed a hypermetabolic left posterior lateral bladder wall thickening consistent with biopsy proven small cell urothelial carcinoma with 2 hypermetabolic left pelvic lymph nodes concerning for metastases	0.558
volberg was treated with neoadjuvant cis etoposide 4 cycles administered from 5 1 17 through 6 30 17 under dr fans care	0.416
pet ct on 6 15 17 following 3rd cycle of chemotherapy showed a persistent focal left bladder wall lesion with slightly increased metabolic activity	0.451

Table 2: Sentiment scores on 10 untrained clinical sentences on model trained with 25 epochs.

Text	Score
he initially presented with gross hematuria in about november and had a ct scan that showed no hydronephrosis but tumor overlying the left posterior wall	0.793
turbt was read as showing high grade tumor with widespread invasion of the lamina propria but no definite detrusor muscle in the specimen	0.440
it sounds like dr nickas was suspicious of muscle invasion	0.333
about 2 weeks after the turbt he had clot retention and was admitted to john muir hospital for a few days and was cared for by dr hopkins requiring catheter placement and irrigation of the bladder	0.482
repeat turbt by dr skinner on 4 10 17 showed a left anterior wall bladder tumor	0.482
resection of tumors of the left anterior wall left lateral wall and tumor base were performed	0.539
pathology showed small cell carcinoma 95 urothelial carcinoma with glandular differentiation 5 invasive of lamina propria and muscularis propria	0.465
staging pet ct on 4 21 17 showed a hypermetabolic left posterior lateral bladder wall thickening consistent with biopsy proven small cell urothelial carcinoma with 2 hypermetabolic left pelvic lymph nodes concerning for metastases	0.434
volberg was treated with neoadjuvant cis etoposide 4 cycles administered from 5 1 17 through 6 30 17 under dr fans care	0.390
pet ct on 6 15 17 following 3rd cycle of chemotherapy showed a persistent focal left bladder wall lesion with slightly increased metabolic activity	0.525

Table 3: Sentiment scores on 10 untrained clinical sentences on model trained with 100 epochs.

3 Future Works

There are many problems with my existing approach.

- The clinical notes data was badly annotated. Due to lack of clinical background, I have really no idea whether the sentences should be perceived as positive or negative. For some where an improvement or no ill effect is noted, I was able to classify it as “positive". However, many of the notes were also simply directions given to the patients, such as important information before surgery or after. For many of those sentences, I ended up randomly assigning a sentiment. Hence, the additional 299 sentence dataset was extremely randomly annotated. This may have resulted in the breaking down of the sentiment analysis algorithm. For the future, I would ask a doctor to properly annotate the clinical notes. Moreover, I should also include the sentiment “neutral" in the analysis so we can label that for the clinical directions.
- The clinical vocabulary was not trained for the word2vec embeddings. Due to simplicity sake, I did not train the word embeddings with the clinical data. However, that also means the clinical phrases and words were not represented in the embeddings. For future, I would include the clinical notes also in the word2vec embeddings training.
- The neural network model was not tuned. For the quick experiment, I did not tune the neural network model and was not very sensitive to the design and structure. The network architecture could make a huge different for future experiments and analysis.

In summary, the experiment showed that sentiment analysis for clinical notes is a long and difficult road. However, NLP analysis of clinical notes is a promising field with many breakthroughs and discoveries.

References

[1] Kerstin Denecke and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27, 2015.

[2] Qufei Chen and Marina Sokolova. Word2vec and doc2vec in unsupervised sentiment analysis of clinical discharge summaries. *arXiv preprint arXiv:1805.00352*, 2018.

[3] Gang Li and Fei Liu. A clustering-based approach on sentiment analysis. In *2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering*, pages 331–337. IEEE, 2010.

[4] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618. ACM, 2013.

[5] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[6] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[7] Javaid Nabi. Machine learning - text processing. `https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958`, Sep 2018.

[8] Javaid Nabi. Machine learning-word embedding sentiment classification using keras. `https://towardsdatascience.com/machine-learning-word-embedding-sentiment-classification-using-keras-b83c28087456`, Oct 2018.